Report No. UT-20.20

A SYNTHESIS OF PASSIVE THIRD-PARTY DATA SETS USED FOR TRANSPORTATION PLANNING

Prepared For:

Utah Department of Transportation
Research & Innovation Division

Final Report
December 2020

RESEARCH

UDOT Keeping Utah Moving

Utah Department of Transportation - Research Division
4501 South 2700 West - P.O. Box 148410 - SLC,UT 84114-8410

DISCLAIMER

The authors alone are responsible for the preparation and accuracy of the information, data, analysis, discussions, recommendations, and conclusions presented herein. The contents do not necessarily reflect the views, opinions, endorsements, or policies of the Utah Department of Transportation or the U.S. Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind, and assumes no liability therefore.

ACKNOWLEDGMENTS

The authors acknowledge the Utah Department of Transportation (UDOT) for funding this research, and the following individuals on the Technical Advisory Committee for helping to guide the research:

- Robert Chamberlin (Consultant, UDOT Research)
- Adam Radel (UDOT)
- Kaitlyn Marousis (UDOT)
- Rikki Sonnen (UDOT)
- Richard Brockmyer (UDOT)

TECHNICAL REPORT ABSTRACT

1. Report No. UT-20.20	2. Government Accession No. N/A	3. Recipient's Catalog No. N/A
4. Title and Subtitle A SYNTHESIS OF PASSIVE THIRD-PARTY DATA SETS USED FOR TRANSPORTATION PLANNING		5. Report Date Dec 2020 6. Performing Organization Code
7. Author(s) Gregory S. Macfarlane, Michael J. Copley		8. Performing Organization Report No.
9. Performing Organization Name and Address Brigham Young University Department of Civil and Environmental Engineering 430 Engineering Building Provo, UT 84602		10. Work Unit No. 5H08282H 11. Contract or Grant No. 20-8511
12. Sponsoring Agency Name and Address Utah Department of Transportation 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410		13. Type of Report & Period Covered Final Nov 2019 to Dec 2020 14. Sponsoring Agency Code PIC No. UT19.602

15. Supplementary Notes
Prepared in cooperation with the Utah Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration

16. Abstract
This report summarizes an extensive review of academic and professional research making use of third-party data products available to transportation agencies. In particular, the review focused on studies conducted with Bluetooth receivers, Global Positioning System (GPS) devices, and mobile device data (MDD) obtained through cellular networks and location-based services embedded in smartphone applications. The review considers the specific applications of the data products seen within the literature, the sample size reported or inferred within the reviewed studies, and summarizes success or failure of the attempted applications. The review also contains a summary of the underlying technologies, including inferred and known sources of bias or incompleteness. The recommendations of this report are based on the previously described review of the literature as well as personal interviews with prominent researchers identified in the review. The central finding of the report is that entirely replacing legacy data collection methods is not likely to result in the best outcomes for UDOT or other agencies. Rather, passive data products can be most useful to agencies that use them as a supplement to existing and ongoing data collection efforts.

17. Key Words Data management, transportation planning, passive data, Bluetooth, GPS, Location-Based Services		18. Distribution Statement Not restricted. Available through: UDOT Research Division 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410 www.udot.utah.gov/go/research		23. Registrant's Seal N/A
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 100	22. Price N/A	

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AADT	Annual Average Daily Traffic
aGPS	Assisted Geographical Positioning System
CDR	Call Data Records
DOT	Department of Transportation
FHWA	Federal Highway Administration
GIS	Graphical Information System
GPS	Global Positioning System
LBS	Location-Based Services
LOS	Level of Service
MAC	Media Access Control
MDD	Mobile Device Data
O-D	Origin-Destination
ODOT	Oregon Department of Transportation
RMSE	Root Mean Square Error
SDK	Software Development Kits
TAZ	Traffic Analysis Zone
TCD	Traffic Control Device
UDOT	Utah Department of Transportation
VMT	Vehicle Miles Traveled
WIM	Weigh-in-Motion

EXECUTIVE SUMMARY

In the last ten years, departments of transportation and other transportation planning and service agencies have begun using passive data products derived from various mobile technologies that reveal the time and location of personal devices and are resold by commercial data vendors for transportation system evaluation and service planning. These data promise to enhance or replace legacy data collection efforts. However, the proliferation of technologies and vendors in this space have made it difficult for transportation agencies to understand both what is available, and what might be possible or feasible with these data.

This report summarizes an extensive review of academic and professional research making use of third-party data products available to transportation agencies. In particular, the review focused on studies conducted with Bluetooth receivers, Global Positioning System (GPS) devices, and mobile device data (MDD) obtained through cellular networks and location-based services embedded in smartphone applications. The review considers the specific applications of the data products seen within the literature, the sample size reported or inferred within the reviewed studies, and summarizes success or failure of the attempted applications. The review also contains a summary of the underlying technologies, including inferred and known sources of bias or incompleteness.

The collective efforts of the many researchers compiled in this document show that each technology has particular strengths and weaknesses. In general, Bluetooth excels in projects of a temporary nature such as construction corridor planning, as well as targeted cordon studies that focus on the entry and exit of individuals within a selected perimeter. Bluetooth can be permanently installed and is often used in real-time applications such as travel-time prediction on highways. Bluetooth falls short in its ability to obtain the true origin and destination of individual trips. GPS technology is highly accurate and precise, and therefore excels in studies that require individual route traces, such as network construction and travel-time studies. Due to the lower penetration of GPS devices across an unbiased population, the technology may be unable to identify wider behavioral information. By contrast, the widespread proliferation of mobile device data (MDD) promise a wider and less biased view of population travel patterns, though the aggregation of data from multiple sources and the coarse temporal resolution of these data limits the precision of studies conducted at small scales.

The recommendations of this report are based on the previously described review of the literature as well as on personal interviews with prominent researchers identified in the review. The central finding of the report is that entirely replacing legacy data collection methods is not likely to result in the best outcomes for UDOT or other agencies. Rather, passive data products can be most useful to agencies that use them as a supplement to existing and ongoing data collection efforts. The passive data can provide scale and context, which can be validated and compared against trusted if relatively expensive active collection methods. It is in the best interest of UDOT to carefully analyze project goals and identify which passive data source may be used to achieve multiple Department purposes. Specific recommendations include:

1. UDOT staff need to be aware of the source technologies of third-party data sets, along with their strengths and weaknesses.
2. UDOT should develop consistent data validation routines to identify the inherent accuracy of purchased data products, and regularly evaluate new data purchases.
3. UDOT should investigate using permanent Bluetooth receivers to measure travel times between points on key corridors. This may be particularly useful in places where GPS and cellular reception might be unreliable, such as in canyons.
4. UDOT should not rely on permanent or temporary Bluetooth receivers to analyze trip origins, destinations, routes, or volumes outside of well-defined cordon studies or institutional settings.
5. UDOT should use GPS data to develop a more complete picture of freight movements in and through the state.
6. UDOT should continue to investigate the use and applications of aggregate GPS data in determining speeds on roadways.
7. UDOT should work to develop an integrated transportation planning approach that relies on the relative strengths of household travel surveys as well as third-party origin-destination data derived from LBS.
8. UDOT should avoid relying on a single vendor or data source for any of its analyses, rather developing processes that make use of multiple data inputs.

CHAPTER 1 - INTRODUCTION

1.1 Problem Statement

In the last decade, a number of departments of transportation and other transportation planning and service firms have begun using data that is collected, aggregated, and resold by commercial third-party firms. These firms have developed business models centered around *passive* data, as the data are passively generated as a result or byproduct of other processes. These processes may include but are not limited to Bluetooth devices searching for receivers to connect with, vehicle drivers navigating with the aid of the Global Positioning System (GPS), cellular phones connecting with network towers, and individual smartphone users using location-based services within their phones. In each of these examples, individual device users reveal their position in space at a defined point in time; on aggregate and with sufficient volume, the collected data can be highly valuable for maintaining and planning transportation services.

The term *passive data* thus stands in contrast to *active* data collection methodologies and systems traditionally operated by departments of transportation. These active systems include traffic counting regimes using manual traffic counts or automatic counters, travel-time studies using probe vehicles, household travel surveys, vehicle intercept surveys, and other ongoing efforts. These data collection and management efforts can be expensive and therefore sparse or infrequent. By leveraging passive data, departments of transportation hope to supplement or even replace these legacy data collection systems.

However, the wide variety in the technologies underlying passive data, and the proliferation of firms aggregating and reselling this data, make it relatively difficult for departments of transportation to evaluate and discriminate between vendors. This task is made more difficult by the opaque sourcing and proprietary algorithms employed by numerous vendors. The Utah Department of Transportation (UDOT) therefore desired a review of academic and professional research employing various types of third-party data.

1.2 Objectives

The purpose of this project is to review the uses of passive third-party data products available to UDOT and other departments of transportation. This includes a discussion of the underlying technologies, how the technologies' specific strengths and limitations lend themselves to studies of different kinds, and a literature review identifying classes of studies that have been attempted with each technology.

1.3 Outline of Report

The report proceeds with one chapter for each technology used by passive data providers. Each chapter includes a description of the technology, a review of studies in the literature that have used that technology, and recommendations to UDOT and other departments of transportation related to that specific technology. The final chapter includes recommendations in the context of the entire report, including a catalog of data vendors and technologies.

1. Introduction: this chapter
2. Bluetooth
3. Global Positioning System (GPS)
4. Mobile Device Data (MDD)
5. Conclusions and Recommendations

CHAPTER 2 - BLUETOOTH

2.1 Overview

Bluetooth is a short-range, radio-based communication protocol that allows pairs of authenticated devices to send limited amounts of data between each other, with one device referred to as the *transmitter* and the other referred to as the *receiver*. Applications of this technology include wireless headphones that connect with mobile telephones, or mobile telephones that connect with in-car audio and information systems. In order for Bluetooth-enabled devices to find each other, transmitter devices constantly broadcast their Media Access Control (MAC) address, an identifier unique to each Bluetooth device. When enabled, Bluetooth devices are continually emitting and searching for Bluetooth signals. Bluetooth receiver units, which are set up by the DOT or its contractors, constantly scan for incoming MAC addresses.

Transportation analysts have exploited this feature of Bluetooth technology to conduct transportation studies. Specialized Bluetooth receivers can be configured to log all devices that ping against them, and the timestamp of each ping. If a device with its Bluetooth functionality engaged passes within the effective radius of a specialized receiver, the receiver can detect the device without line of sight and through solid objects. By tracing which devices ping against a set of spatially dispersed receivers, transportation analysts can reconstruct travel times between receivers, and route choices through a network.

This chapter begins with an overview of Bluetooth technology that informs a subsequent review of transportation studies conducted using Bluetooth devices. We then synthesize strengths and weaknesses of using Bluetooth data, or scenarios in which it would be appropriate or not based on our observations and lessons learned from the literature review.

2.2 Bluetooth Technology

Bluetooth originated in the late 1990s and has been under continuous development ever since (Bluetooth SIG, 2020). Bluetooth 1.0 was released in 1999 with a maximum range of 33 feet which was improved to 200 feet in version 2.0. By 2011, Bluetooth 4.0 could reach a range of 200 feet and version 5.0, released in 2016 was able to reach 800 feet (“Bluetooth,” n.d; Triggs, 2018). As of August 2020, Bluetooth 5.2 is the newest version of Bluetooth software available on the market. Bluetooth has improved in its range, data transfer rates, and other features as a function of advances in both signal processing technology and electrical device manufacturing and design. Previous to version 5.1, only the approximate location of a transmitter could be determined from within a receiver's range radius. Bluetooth 5.1 introduced algorithms that enable the receiver to determine the angle of arrival and angle of departure for signals to and from a transmitter device, potentially locating a transmitter to within inches (Hollander, 2019).

Currently Bluetooth 5.0+ has an extreme maximum transmission range of around 780 ft based on the signal processing technology used, and individual Bluetooth devices may or may not have the power necessary to transmit over the maximum range. Despite the advancements of Bluetooth range, each device is limited by its particular sub-class. There are three primary classes of Bluetooth devices with varied ranges. Class 1 has an expected range of 333 ft. Class 2 – which is most common in wireless headsets and smartphones – has a range of 33 ft. Lastly, Class 3 has a 3-foot range (Samsung, 2018; Rescot, 2011). Zinner (2012) indicates that class 3 devices are usually not detected because of their short range. The class of a device is important because Bluetooth receivers installed by DOTs require a certain range in order to detect a device at all. There is also a tradeoff between data transfer rate and range. An increased range would mean a lower data transfer rate which could be a limiting factor for maximum Bluetooth range (Cross, 2018; "Bluetooth," n.d.).

Bluetooth's low power consumption and relative security has led to its use as a prominent technology in consumer devices, with the average American owning approximately 4.4 Bluetooth devices, and 5.5 billion devices are projected to ship to the United States in 2021 (Karr, 2016). Bluetooth in highway vehicles is found not only in occupant devices, such as smartphones and in-car entertainment systems, but also in vehicle status and operation equipment. Modern tire-pressure monitoring systems and anti-lock brakes communicate with computers inside a car using Bluetooth (Vasantharaj & Krishnamoorthy, 2016). This means that Bluetooth receivers on the side of the road can collect Bluetooth MAC addresses from vehicles even at times when user devices are not engaged. It also means that a receiver device may record multiple unique device hits from a single vehicle.

2.3 Transportation Studies Using Bluetooth Data

Transportation analysts can install Bluetooth receiver units and configure them to log the time when a transmitter device passes through the receiver's detection area. This combination of unique device position and time information has been useful to transportation analysts in several different kinds of studies, which we have grouped into three general applications:

1. Travel-Time Studies
2. Cordon Analysis
3. Path and / or Activity Analysis

Schematics of the Bluetooth device configuration necessary to conduct each type of study are given in Figure 1. As shown, the requisite number and density of Bluetooth receivers increases with each study.

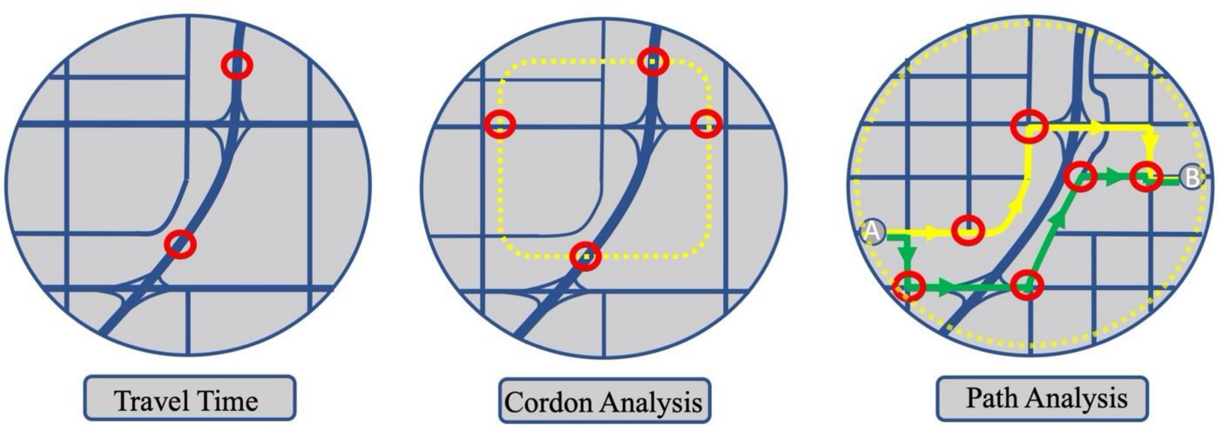

Figure 1. Diagrams showing the different receiver arrangements for various studies.

2.3.1 Travel-Time Studies

Cities and state departments of transportation can use Bluetooth on roadways to measure travel times. These systems are widely used in North America, and services from many companies exist. This section will describe how travel times have been used in various applications. Figure 2 depicts a simple travel-time setup on a highway. To obtain travel times, a minimum of two receivers must be deployed and a Bluetooth device must be detected at two or more separate receiver units.

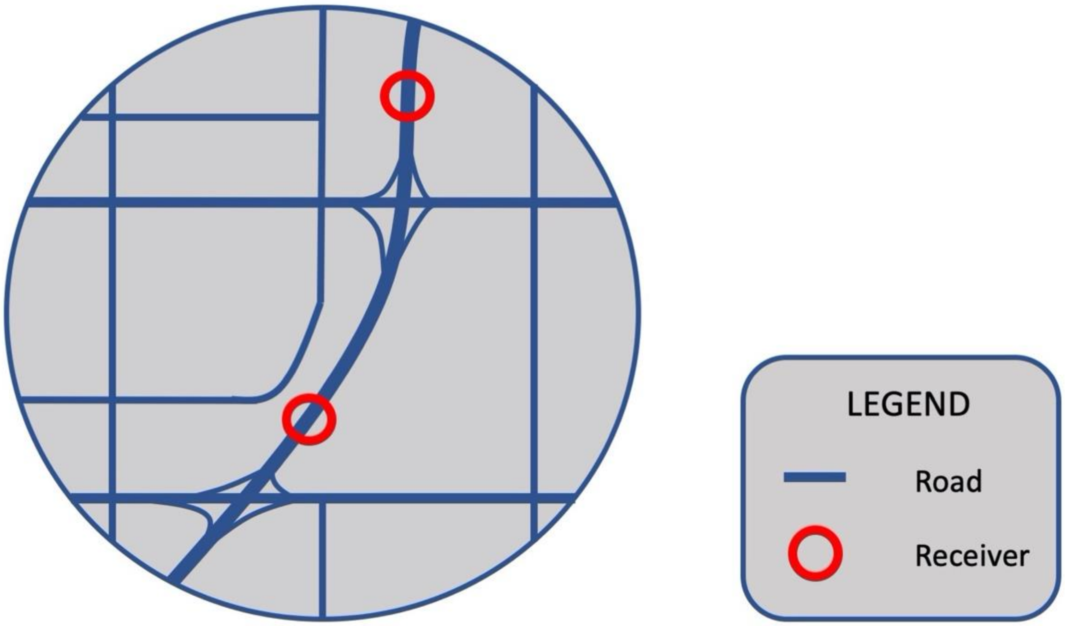

Figure 2. Diagram showing a basic travel-time setup.

Pairs of Bluetooth receivers installed on highway segments can determine travel times of passing vehicles. In Oregon, Porter (2011) tested various Bluetooth receivers on different segments of highway. In their various experiments they discovered a tradeoff existed between timestamp accuracy and penetration rates. Penetration rates refer to the number of vehicles detected out of the entire population. See Table 1 for a list of penetration rates of all major Bluetooth studies in this literature review. Porter (2011) determined that individual circumstances may require variations of receiver units to obtain accurate travel times. In Australia, Blogg et al. (2010) set up receivers at intersections surrounding a major freeway. Throughout the study, the researchers reported long queues at traffic signals before traffic could enter the freeway. This research also found consistently higher penetration rates when compared to other studies. It was observed that the longer an individual resided within a Bluetooth receiver radius, the more likely a Bluetooth device would be detected. This study outlines the strong correlation between setup of receivers and higher penetration rates. They confirmed that Bluetooth is a reliable method of obtaining travel times on highways. In Kansas, Rescot (2011) compared Bluetooth travel times to a GPS-equipped probe vehicle. The researcher deployed 10 portable receivers on a highway segment to identify Bluetooth travel times. Data collected from a probe vehicle traveling at free-flow speed was used as a comparison to the Bluetooth receiver data. Travel times on a 6.5 mile stretch of highway were on average 32 seconds greater with Bluetooth data than with GPS probe vehicle data, but they still determined that Bluetooth was a viable performance measure for identifying travel times on freeways. Both data sources were collected before and after an upgrade was made to a traffic signal's hardware. It was noted that the portable Bluetooth receivers were susceptible to battery failure and subsequent data outages. On several occasions, there were no Bluetooth devices detected for multiple hours. The lack of Bluetooth detections over several hours prevented most comparisons with the GPS probe vehicle, but when comparison was made to the GPS probe vehicle data, it was noted that Bluetooth data consistently had longer travel times which indicates slower movements through the corridor. Also, the Bluetooth data indicated consistently longer travel times after changes to the intersection. From these findings it is recommended that further research be conducted into determining why travel times are longer for Bluetooth data.

Accurate travel-time data can be put into dynamic traffic models and used to forecast travel times. Near Barcelona, Spain, Barceló et al. (2010) confirmed the reliability of travel-time estimations from Bluetooth MAC timestamps. Like most studies, when a unique MAC address is detected by a receiver at two separate locations, an average speed between those two points can be estimated. To detect future travel times, the researchers input mixed historical travel-time records and real-time Bluetooth speeds into a multi-step algorithm. The researchers concluded that forecasted travel times closely resembled the actual travel times. This paper confirms that travel-time forecasts can be determined from a mix of historical data and real-time Bluetooth data.

In a similar study, Haseman et al. (2010) used travel times collected by Bluetooth receivers to predict delay in a construction zone. Travel times were collected at the beginning and end of the work zone. Data was used in real-time to estimate flow of traffic and subsequent delays. The delay estimation was sent to automatic traffic messaging signs to warn upstream traffic. They found that early warnings had little effect on alternative routes selected by drivers. If message boards instructed individuals to select an alternative route, early warnings had a greater effect. Analysis revealed a correlation between accidents and increased delay time within work zones.

Organizations can use real-time trip data to mitigate congestion on freeways during construction, rush hours, natural disasters, and emergencies such as car accidents.

Bluetooth can be used as a validation tool for signal coordination changes. Results from manual counts, which are traditionally used as validation, can be expensive and slow to obtain. Bluetooth can collect data accurately, quickly, and economically, after adjustments are made on any corridor. Kim et al. (2014) used receivers to detect travel times before and after changes were made to traffic signals. Day et al. (2010) used Bluetooth travel times to validate changes made for signal coordination. After analyzing collected Bluetooth data, the researchers confirmed that commuters saved an average of 1.9 minutes from the signal coordination adjustments. This analysis tool was validated using manual counts. These studies provide evidence that Bluetooth can be used as a supplemental analysis tool to identify before-and-after changes to signal timing changes.

To measure the effectiveness of traffic control devices (TCDs), Kim et al. (2014) set up receivers at a three-way stop T-intersection. The aim of this study was to determine how TCDs affect travel time through a region. Kim et al. (2014) noted that travel times were often underestimated by an average of 1.2 seconds. This information could be used as performance measures for analyzing the effectiveness of TCDs. Like the Day et al. (2010) study, reducing the need for manual counts could reduce error, costs, and time to complete an analysis.

Emphasis on active transportation makes mode differentiation increasingly important to agencies. Knowing the movements of cyclists and pedestrians allows agencies to target future projects as the needs of cyclists and pedestrians differ from automobiles. In many instances the needs for separate bike lanes and convenient pedestrian walkways are of lower priority than automobile projects. Active modes may mitigate environmental concerns and possibly help with urban congestion. Differentiated data could identify where improvements are needed and reveal future projects that will encourage the use of active modes. Bathaee et al. (2018) and Kim et al. (2014) used travel times to differentiate modes. These two papers found that Bluetooth could accurately differentiate between vehicles, pedestrians, and cyclists in most conditions. These related papers used clustering methods such as grouping behaviors that were similar for a particular mode. Their work excluded accidents and weather events. Further research should be conducted to include unaccounted-for factors such as dense urban environments where signalized intersections may blur mode choice because speeds are similar between modes. The inherent issue with basing mode differentiation on travel times occurs when traffic signals or congestion are present. In these situations, bicycles and pedestrians could travel quicker than a vehicle. The current algorithm does not account for these urban conditions. Findings indicate that other methods exist such as StreetLight Data (2019) that use mobile device data combined with machine learning techniques in place of travel times. Further research should be done to determine the best methods for mode differentiation.

2.3.2 Cordon Analysis

Several studies have conducted research to prove the effectiveness of using Bluetooth cordon techniques to find Origin-Destination (O-D) information. Although this is not a true O-D as it lacks the actual starting and ending points of a majority of individuals, it does provide entry and exit tracking data from within a specified cordon or perimeter. Figure 3 represents a typical cordon study which has an outer perimeter with receivers at entry and exit points along that perimeter. In this example, not all entry and exit points have a receiver which will result in missed vehicle detections. This paper uses the term “bleed-off” to refer to when a device exits or enters through an unmonitored point.

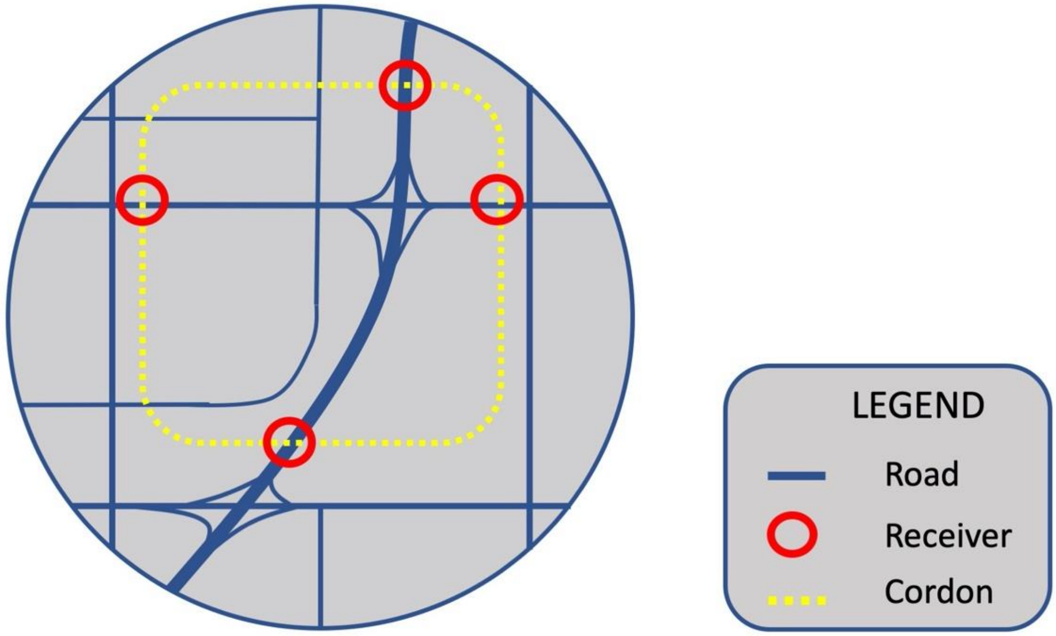

Figure 3. A basic cordon with four Bluetooth receivers.

Bluetooth can be used to obtain O-D information. In Oregon, Kim et al. (2014) successfully obtained around 6% penetration rate on a rural cordon. Several receivers were deployed on major roads surrounding a rural town. By deploying multiple receivers at one location, they found that Bluetooth receivers often failed to detect Bluetooth devices. A receiver directly across a wide road often did not detect a vehicle when the adjacent receiver did. This highlights the need for agencies to carefully install receivers on wider roads. They were successful in obtaining O-D information that could be interpolated to obtain traffic count estimations. The author did not specify if they were successful in obtaining counts or volumes, but they were successful in obtaining O-D information.

Bluetooth can track the O-D of traffic on and through highway networks. Barceló et al. (2010) found mixed results on a cordon study surrounding a highway. Each on and off ramp was fitted with a receiver to track the vehicle location along a highway in Spain. The researcher successfully implemented algorithms to determine travel-time forecasts. When the researcher tried to apply the same method to a cordon, they found mixed results. In free-flow traffic the method worked, but congestion caused the method to fail. They concluded that future research was needed to refine their method for use on highway cordon studies. Chitturi et al. (2014) used aerial photography time lapse to validate Bluetooth data on a clover leaf overpass. Although slightly less accurate, they found that Bluetooth was a reliable method for tracking O-D pairs. Similar to Barceló et al. (2010), each exit and entry point was fitted with a receiver. The researchers discovered that penetration rates from Bluetooth data were generally lower than photographic methods, but Bluetooth was cheaper and provided similar results. It was also noted that studies using temporary Bluetooth installations could easily be lengthened with little or no additional cost. If the study is reliant on manual counts the study may not be lengthened due to budget constraints. Bluetooth receivers, on the other hand, simply need to be left out for longer periods of time, if needed. These studies have identified that Bluetooth can be used to track the O-D of vehicles through a highway network.

Bluetooth has been used to estimate traffic counts at traffic circles. A case study in Rescot (2011) described a traffic count study at two traffic circles. One traffic circle had four approach lanes while the other had five. A Bluetooth receiver was placed at each approach to identify O-D as well as traffic counts. The results of Bluetooth traffic counts were inconclusive after being compared with manual traffic counts. At one traffic circle, results were deemed acceptable and the other traffic circle had opposing results. Although traffic counts were inconclusive, the researcher was able to obtain accurate O-D data from the cordon study. Further research should be conducted to see the effectiveness of a traffic count study at a roundabout.

Obtaining current traffic patterns can aid in predicting the impact of a proposed development. In Turkey, Yucel et al. (2012) conducted a cordon study surrounding a proposed hospital site. A large vacant lot was surrounded by a busy road with many intersections. The researchers set up a partial cordon that covered approximately half the circle around the vacant lot. They failed to produce a reliable cordon as there was a low penetration of O-D pairs. This is likely caused by high bleed-off to side roads and the failure of receivers to re-detect MAC addresses. This study outlines the importance of the correct placement of receivers, and of having a receiver at every possible entry and exit point.

2.3.3 Path Analysis

At times it is beneficial to understand the unique traces through a cordon perimeter. Not only is an outer perimeter set up, but receivers are placed within the cordon to track exact movements within a region. In some instances, pedestrians are tracked, and duration of stay can be determined by timestamps collected within the cordon. Figure 4 depicts a path analysis from specified points residing on the cordon perimeter. In a path analysis study, we are concerned about how a device travels within a cordon as well as when a device enters and exits the system.

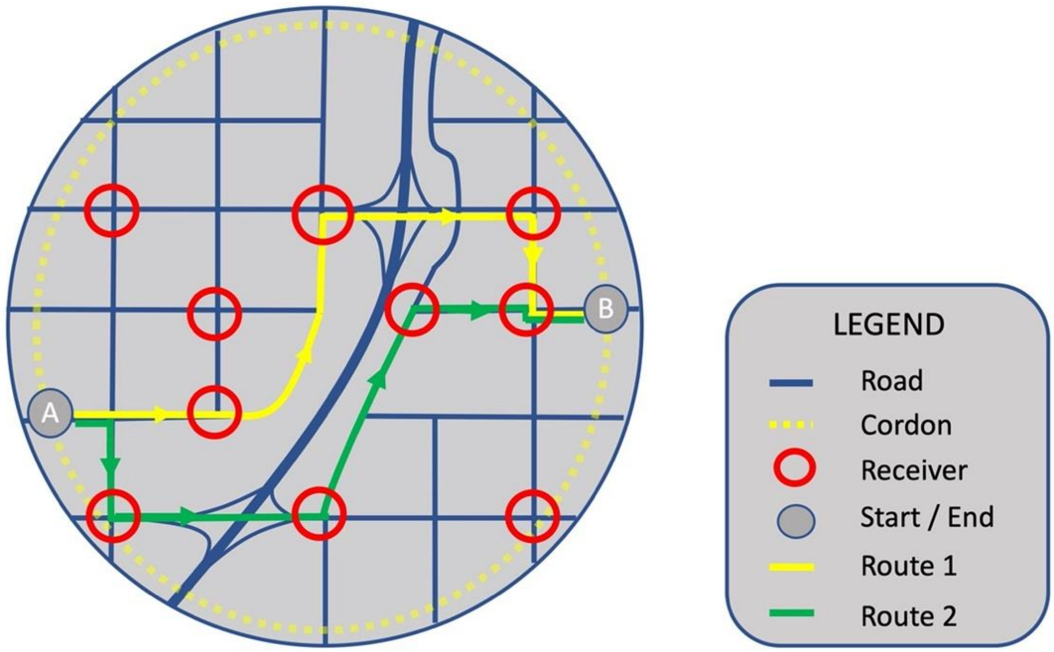

Figure 4. A setup with receivers located inside a cordon to determine specific traces.

For one-day events that temporarily increase local traffic, it can be beneficial to understand the paths chosen to exit a region. In another case study, Rescot (2011) investigated how the University of Missouri could mitigate congestion following significant sporting events. Multiple highways surround the university and event planners wanted to know what percentage of traffic went north, east, and south. Starting at the sports complex, twelve Bluetooth receivers were set up to form a partial 2.5-mile radius perimeter around the University. A smaller perimeter around the sports complex was also set up. Travel times and route choice could be determined from observing traffic dissipate from the smaller perimeter to the outer perimeter. The setup of two perimeters allowed researchers to determine the percentage of fans who went in each direction. Video surveillance was used to validate license plates and confirm O-D pairs. This experiment had high bleed-off because of the high number of alternative routes available. The study obtained a total average penetration rate of 6%. Although travel time and O-D identification closely resembled traditional methods, it was noted that the processing of Bluetooth data was significantly easier than organizing manual counts. The study could be extended without increasing the effort to process data. This study could help with mitigation strategies when single-day events occur. Understanding where the majority of individuals will travel can allow for better signal coordination and other congestion mitigation strategies.

Complex analysis of routes through a cordon can be obtained using Bluetooth data. Jackson and Dichev (2014d) also used video surveillance as a validation source and found that Bluetooth cordon data closely resembled the more costly automatic number-plate recognition technique. This study used one of the most extensive Bluetooth cordons to date. One hundred and ten devices were deployed to encircle the London congestion charge zone. The researchers set up two external loops to better capture all vehicles for their path analysis study. The use of 110 receivers resulted in one of the highest penetration rates in this literature review and was able to identify common travel patterns and volume estimations. Not having enough processing power to store large volumes of Bluetooth data was a recurring issue. They also developed their own adjustment formulas to remove bias from multiple Bluetooth devices detected in one vehicle. The authors' method for detecting multiple devices in one vehicle was not specified. In a similar study Blogg et al. (2010) used 29 receivers in Brisbane, Australia to identify route choice through an arterial network. Approximately 15% of the population was detected and expansion factors were used to match results to the entire population. Results for this study were inconclusive as they determined further research was necessary to confirm the use of MAC addresses in determining O-D pairs. They noted that results compared favorably with automatic number-plate recognition, but several possible biases may exist that need further investigation. Carpenter et al. (2012) had success using Bluetooth data to analyze specific routes chosen on a focused corridor. In a rural area outside Jacksonville, Florida, Bluetooth receivers were placed on a major corridor and its crossroads. The many crossroads provided several alternative routes. After data was filtered the researchers successfully derived travel time, use percentages (road use rankings), and O-D pairs. They concluded that Bluetooth was a valid tool when determining specific routes traveled through a cordon. They suggested that this technique be used for traffic forecasts and in traffic model validations.

Attempts have also been made to track free movements of pedestrians using Bluetooth proximity techniques. Oosterlinck et al. (2017) used Bluetooth and Wi-Fi to track pedestrian movements in a shopping mall and were able to aggregate data to find common patterns of overlap. Utsch et al. (2012) discussed how Bluetooth can isolate what they refer to as microscopic movements and habits of individuals. The researchers set up predefined paths and then used Bluetooth proximity to match pedestrians within the predefined squares. From these pixel-like grids, generalized paths could be formed. Isolating such patterns had the ability to map foot traffic patterns. Since pedestrians are not bounded by lane restrictions, one can isolate exact path trends of individuals. Bluetooth can also help plan logistics of events such as determining adequate restroom facilities, hallway width, and advert placements based on quantity of detections. Combined quantity and duration statistics can help to identify bottlenecks within a system.

Duration of stay can be derived from multiple readings at one or more Bluetooth devices within a venue. A startup company in Salt Lake City has used this technology to track attendance and patronage at the Sundance Film Festival. This information can consequently be used to plan logistics of parking, venue size, and public transit. O'Neil et al. (2006) found that using Bluetooth to track dwell times of customers in a coffee shop was not a standalone method. Supplements from video surveillance were used to accurately determine patron dwell time. Kurkcu et al. (2017) discusses research done to estimate wait times at bus stops/terminals using Bluetooth and Wi-Fi. The results were successful in finding first and last scans for unique MAC address, which could be used as a performance measure to decrease wait times. The method could also be used to estimate pedestrian densities and flow. Further research should be done to see how Bluetooth can be used to monitor pedestrians.

2.4 Analysis of Bluetooth as a Data Source

This section will discuss the strengths and weaknesses of Bluetooth technology so that agencies will understand how to best implement the technology. The section will first discuss trends in penetration rate and then will summarize strengths and weaknesses.

2.4.1 Penetration Rates

Penetration rate refers to the number of vehicles detected out of the entire population. It is important to understand what the penetration rate is so that data can be expanded to represent entire populations. This section will discuss findings on penetration rate.

Table 1 summarizes various studies and the penetration rates recorded from that study. Each study was deemed successful based on each researcher's opinion and observations, and a summary table was created to observe general trends in a limited number of studies. Penetration rates as high as 40% are achieved but regularly average 6-10%. Lower penetration rates are sufficient for travel times but may not provide enough detail for cordon studies (especially real-time applications). Notice that penetration rates in the figure represent an average penetration rate collected at a minimum of one Bluetooth receiver. The chart is organized from lowest to highest penetration rates. This compilation of studies suggests no correlation between increasing penetration rates over time. The data suggests that more receivers correspond to higher cordon penetration rates. Notice Blogg et al. (2010) and Jackson and Dichev (2013) both have larger numbers of receivers and corresponding higher penetration rates.

Table 1. Penetration Rates Observed in Bluetooth Studies

Author	Year	Penetration (average %)	# of Receivers	Study Region (miles)	Success	Notes
Travel Time				***Corridor Length***		
Kim et al.	2014	6.3%	3	200 (feet)	Yes	Intersection - TCD study.
Kim et al.	2014	6.6% Cars – 6.8% Bicycles - 40% Pedestrian - 10.3%	8	200-400 (feet)	Yes	Mode Differentiation / Intersection Travel time /
Bathaee et al.	2018	Cars - 6.8% Bicycles - 14.5% Pedestrian - 7.2%	8	400 (feet)	Yes	Mode Differentiation
Haseman et al.	2010	8%	9	10.7	Yes	Construction Zone
Porter et al.	2011	5.7 - 9.6%	2-5	1-6	Yes	Travel Time
Blogg et al.	2010	38%	6	8-15	Yes	Travel Time
Day et al.	2010	-	9	3.2	Yes	Signal Timing
Barceló et al.	2010	-	6	7-25	Yes	Travel Time
Rescot	2011	-	10	6.5	Yes	Highway
Rescot	2011	-	10	2	?	Arterial Signal Adjustments
Kim et al.	2014	-	3	1.5	Yes	Signal Timing
Cordon				***Cordon Radius***		
Yucel et al.	2012	2.8 - 4.1%	4	1	No	Inner City
Chitturi et al.	2014	4.4%	5	0.5	Yes	Overpass
Kim et al.	2014	6.6%, 6.2%	3, 13	25	Yes	Rural / Multi City
Rescot	2011	6-8%	4-5	100 - 330 (feet)	O-D Yes Counts, No	Roundabout, Traffic Counts
Barceló et al.	2011	-	6	7-25	Free flow, Yes Congestion, No	Highway
Path Analysis				***Cordon Radius***		
Carpenter et al.	2012	6.1%	14	15	Yes	Route Choice
Rescot	2011	6%		2.5	Yes	Route Choice
Oosterlinck et al.	2017	9.8%	56	1	Yes	Indoor, Ped.
Blogg et al.	2010	20.0%	29	8-15	Yes, if used as Supplemental Data	Inner City
Jackson & Dichev	2013	30.0%	110	12-15	Yes	Inner City
Rescot et al.	2011	-	10	5	No	Route Choice

Penetration rates of Bluetooth data are not constant. The penetration rate changes with time and exhibits random behavior (Barceló, 2010). Rescot (2011) explored how capture rates vary. Bluetooth proved to be viable, but the availability of a consistent stream of Bluetooth devices was not. For instance, even on a freeway there is a chance that no devices are detected in a given time. No detection could be caused by low traffic volumes or the possibility that no one has a Bluetooth device in discoverable mode. Even during peak times on low-volume roads there is evidence that certain time intervals will have 0% detection. To counter such problems, a longer sample time is recommended to ensure that these outliers are neutralized.

Several questions regarding detection and bias in Bluetooth technology exist. For instance, data may be skewed towards newer vehicles which have ABS Bluetooth detectors or towards users who have expensive cell phones. These potential biases may skew data to those with higher incomes. Applying expansion factors to the potentially skewed sample for the purpose of estimating the overall population has the potential of amplifying the bias in that sample. Multiple devices in one vehicle can also skew results. Jackson and Dichev (2013) developed a filter that could detect if a vehicle had multiple devices, and then could record data from only one of the devices in that vehicle. Little research has been done to address these potential biases and future research should be conducted.

Travel time can be derived from low sample rates. Previous findings indicate that, for travel-time applications, lower penetration rates tend to mimic real-life traffic conditions well for most purposes with added benefits of smaller data sets which take less storage. One interviewed company claims that their penetration rate is as high as 40% on average (Anonymous, personal communication, November 7, 2019). A UDOT employee countered by summarizing that high penetration rates are unnecessary because low rates have similar distribution curves to high penetration rates and therefore both data sets are usable for a department's purposes (G. Farnsworth, personal communication, November 2019). Table 1 shows that penetration rates typically reside between 2% and 12%. When using Bluetooth to detect travel speeds in congested urban areas, lower penetration rates are still useful because one may assume that traffic flow will be more consistent due to higher congestion. Thus, generalizations can be made from lower penetration rates. Porter et al. (2011) states that a low penetration rate of 2-3% is sufficient to collect travel-time estimations, but higher rates may be needed on arterial roads. Table 1 indicates that Bluetooth penetrations as low as 5% seem sufficient to derive traffic travel times and cordon enter/exit points, though a 5% sample may not be sufficient for complex cordon geometries. Minnesota Department of Transportation found that Bluetooth was as effective as their Automatic Traffic Recorder systems in measuring average travel speeds, and that penetration rates were dependent on equipment setup (Young, 2012).

High penetration rates have been obtained in some studies. Blogg et al. (2010) discovered rates as high as 40%. The researchers pointed to large queue lengths at highway onramps as the primary reason for the abnormally high rates. They mentioned that the longer a vehicle resided within a receiver's radius, the more likely a Bluetooth device would be detected. It was also discovered that penetration rates increased as much as 7% when multiple Bluetooth receivers were installed at each entry point into a cordon. This study may suggest that there is a difference in penetration rates with free-flow traffic versus congested traffic. Further research into this topic may be of benefit. From Table 1, a loose correlation can be observed which suggests that a higher number of Bluetooth receivers may result in higher penetration rates. Not only the number of receivers, but the correct placement and distribution of the receivers would likely influence penetration rates. These interpretations should be investigated further.

2.4.2 Strengths

Bluetooth technology allows agencies to target specific areas and customize data collection. Bluetooth receivers can be installed permanently or deployed as temporary battery-powered units. Setup of portable units is generally quick, simple, and inexpensive which allows for easy adjustments. Maintenance for the portable units is also generally simple. Another advantage of Bluetooth is its ability to detect devices without line of sight and through solid objects, though not every device will be detected within range.

The cost of Bluetooth is relatively low. Chitturi et al. (2014) noted how studies could be easily lengthened at little or no extra cost to agencies because devices could be left out. In the past, traditional video or aerial methodologies would require significant manpower to obtain data for analysis. Expenses from employee wages as well as the delay in obtaining reliable data could cost agencies small fortunes. Rescot (2011), and Day et al. (2010) both indicated that the use of Bluetooth for analysis could save significant money. Haseman et al. (2010) and Porter et al. (2011) indicated how simple and cost effective it is to set up a temporary cordon study. Bluetooth units are relatively inexpensive and easy to maintain. In addition, the setup can be adjusted easily as changes are needed.

Data can be obtained and used in real time with Bluetooth receivers. Bluetooth data is relatively straightforward compared to other passive data sets which may require significant processing or aggregation. Haseman et al. (2010) used Bluetooth in a construction zone to detect travel delays. Information can be collected and transmitted to automated messaging signs or online traffic forecasting tools. Real-time speed information could also be used for dynamic speed limits, ramp metering, etc.

Fewer privacy issues exist with the use of Bluetooth data. Bluetooth data does not collect the actual starting and ending points of trips unless an individual begins and ends within a cordon. The lack of trip purpose and, as the unique identifiers of a Bluetooth device are rarely registered to a specific individual, identification of unique persons is challenging. Of the passive data sources discussed in this report, Bluetooth has relatively low privacy concerns.

Bluetooth devices are increasingly available. Technology changes rapidly which may deter agencies from investing long term in technology that may quickly become obsolete. Bluetooth devices have become more common in consumer products. As of 2016, the average American owned 4.4 Bluetooth devices, and a projected 5.5 billion devices were estimated to ship in 2021 (Karr, 2016). The popularity of Bluetooth in consumer devices provides a longevity that makes the receiver technology worth investing in.

Travel time can be derived from low sample rates. Lower penetration rates tend to mimic real-life traffic conditions well for most purposes. High penetration rates are unnecessary because low rates have similar distribution curves to high penetration rates and therefore both data sets are usable for a department's purposes (G. Farnsworth, personal communication, November 2019). Low Bluetooth penetration rates can be used to identify travel times.

Bluetooth technology works well in situations of a temporary nature such as construction sites, single-day events, or post-project analysis. The processing of Bluetooth data is significantly easier than organizing manual counts. Rescot (2011) successfully used Bluetooth data to mitigate congestion following significant sporting events at a Missouri University. Haseman et al. (2010) used Bluetooth receivers to help during the renovation of a highway. The inexpensive and easy setup of Bluetooth receivers allows for temporary operations that can easily be taken away once an event or project is completed.

2.4.3 Weaknesses

To collect Bluetooth data, a DOT or their contracted vendor must set up and maintain a network of Bluetooth receivers. Data is only collected in areas where receivers are set up. Unlike GPS or MDD data sources, Bluetooth data is not continuously generated without a receiver set up to record timestamps of unique MAC addresses. As a consequence, many gaps in historical data usually exist which makes the data better suited for real-time applications or future projects. In some cases, companies that specialize in Bluetooth data may have previously collected data sets. Although Bluetooth is often used in real-time applications, best results are obtained when data is aggregated over longer time periods because of the nature of sampled data (Rescot, 2011). Third-party reliance can become an issue as Bluetooth equipment is not generally transferable to another vendor. Future improvements to infrastructure would require partnership with the same vendor. It is possible to collect data as a DOT with portable units for a smaller study, but larger projects may require a third-party vendor's expertise.

True O-D information is not collected. A cordon must be created which lacks details of where individuals are starting and ending their trips. Bluetooth can only determine when an individual entered and exited the customized perimeter. The lack of trip purpose limits how Bluetooth data can be used. Installing receivers at every side street, business, and alleyway is often impractical. If every entry and exit point cannot be fitted with a receiver then bleed-off will occur.

Portable Bluetooth receivers are more susceptible to lag and battery failure than probe-based data sources such as GPS and MDD. Bluetooth receivers can be installed as permanent units or set up temporarily on a job site using portable battery-powered devices. Rescot (2011), who was using portable units, mentioned the failure of a receiver unit which caused data outages that damaged their experiment. A permanent receiver installation may also be susceptible to power outages where traffic signals become non-functional. Data outages can also occur when receivers have a minimum threshold to record travel times. In a busy Utah canyon, where an unexpected traffic jam occurred, governing algorithms caused receivers to stop collecting data when traffic slowed below the speed threshold for data collection. Minimum speeds in the algorithm were set lower to avoid future problems. Lowering the minimum detection speed in the canyon, which typically has highway speeds, allowed bicycles to be detected which UDOT did not want to include (UDOT, personal communication, March 3, 2020).

It is possible to use Bluetooth for mode differentiation in free-flow conditions, but Bluetooth data cannot be used to differentiate modes on arterials. Bathaee et al. (2018) discovered a method that could successfully discriminate between modes. They conclude that discriminating between vehicles, cyclists, and pedestrians is 100% effective in ideal circumstances. Further research was needed to account for accidents and weather events. Oregon Department of Transportation (ODOT) has also done research on mode differentiation and found that Bluetooth was sufficient. Their clustering approach had similar results to the previously mentioned study. They found that mode differentiation was increasingly difficult near signalized intersections as the behavior of individual modes was similar and could be confused by their algorithms. They suggest additional research be done in this area (Kim, 2014). Both of these studies used travel times to distinguish between modes. In urban environments a cyclist or pedestrian may travel faster than a vehicle which may render such methods ineffective. In most places where one would want modes differentiated, the existing algorithms are unable to provide reliable results.

Sample sizes of Bluetooth data are not constant. The sample size changes with time and exhibits random behavior (Barceló, 2010). Rescot (2011) explored how variable capture rates were. Bluetooth proved to be viable but the availability of a consistent stream of Bluetooth devices was not. For instance, even on a freeway there is a chance that no devices are detected. No detection could be caused by low traffic volumes and the possibility that no one has a Bluetooth device in discoverable mode. Even during peak times on low-volume roads there is evidence that certain intervals will have 0% detection. To counter such problems a longer sample time is recommended to ensure that these outliers are neutralized.

Obtaining travel times and O-D requires a unique MAC address be identified by two receivers which is sometimes challenging (Yucel, 2012). Often a MAC address is only detected once, which makes the data point largely unusable. For example, to obtain a travel time, a reading of the same MAC address at two different receivers is necessary. Kim et al. (2014) found that in a controlled environment, 19 out of 20 trial runs detected a unique MAC address at the two test receivers. In uncontrolled environments the recapture rate at a second receiver varies significantly. In Yucel et al. (2012), 9% - 14% penetration was captured at one receiver while only 2.8% - 4% penetration was obtained at two locations. Chitturi et al. (2014) found similar results where 4.4% penetration rate was obtained at one receiver while 2.3% penetration was obtained from two receivers. Blogg et al. (2010) had 20% penetration at one receiver and 15% at two or more receivers. The minimum threshold for obtaining MAC addresses in two or more locations appears to be around 5% of the total population (Tao, 2012). Occasionally Bluetooth devices, which are within a receiver's detectable radius, are not detected. This phenomenon has little explanation and generally has a low impact on the data collection process. Identifying a unique MAC address at two or more different receivers is sometimes challenging but is required for any useful transportation application.

Some argue that Bluetooth sensors struggle to select when to time stamp vehicles at signalized intersections. When timestamps are premature the predicted travel time is often longer than the actual time. This may be caused when a vehicle is waiting for a signal to turn green. An Oregon DOT employee in charge of regional Bluetooth data revealed the difficulty in obtaining proper timestamps for arterial networks, especially for vehicles caught at red lights (R. Gamble, personal communication, October 3, 2019). Vehicles within urban settings typically showed longer travel times, and it is hypothesized that the premature timestamping of Bluetooth-enabled devices skews results toward longer estimated travel times. Rescot (2011) suggests that bias exists from distracted drivers on their cell phones, which cause the driver to move slower through a corridor, hence longer travel times. Opposite to longer travel times, Kim et al. (2014) found that travel times were often underestimated with an average of 1.2 seconds under true travel time. Porter et al. (2011) discovered the angle of a receiver antenna affected when a timestamp was made and how early a Bluetooth device was detected. The researchers main challenge, which still exists today, is how to filter out multiple timestamps within one area and then to time stamp the individual when it is closest to the receiver. Timestamp problems also affect travel times when a vehicle is obstructed by red lights, accidents, etc. It appears that there is no precise method for choosing the best timestamp. Zinner (2012) mentions that Class 1, 2, and 3 Bluetooth devices all transmit at different speeds which will affect when the timestamp occurs. Class 1 devices, which are the most powerful Bluetooth strength devices, have the fastest transmission speeds. Also, Class 1 devices have larger range which could be detected by multiple receivers if receivers are placed closely together. Blogg et al. (2010) indicated that the longer a device is within a receiver's range, the more likely it is to be detected. The same researcher also found that having a second receiver at every location increased penetration rates up to 7% more. Additional research should be conducted to determine if newer versions of Bluetooth would affect when a timestamp occurs. Although time differences are usually small, further research should be conducted on appropriate timestamping methods and causes for the common errors with travel times. Porter et al. (2011) discovered that a tradeoff exists between obtaining accurate timestamps and penetration rates. After testing various Bluetooth receivers on different segments of highway, it was determined that higher penetration rates accompanied less accurate timestamps and more accurate timestamps had lower penetration rates.

2.5 Summary and Recommendations

This chapter explained how Bluetooth technology functions, how transportation agencies can use the technology, and provided an analysis of strengths and weaknesses. Bluetooth is a short-range, radio-based technology that allows authenticated devices to send limited amounts of data between each other. Bluetooth is a relatively inexpensive technology that has relatively high privacy for the information collected. The inclusion of Bluetooth technology in consumer devices has increased over the past decade and provides an opportunity for passive data collection. Findings indicate that Bluetooth provides accurate results that closely replicate traditional methods such as automatic number-plate recognition and traffic surveys. Bluetooth has the ability to transmit a unique MAC address that can be captured at multiple receiver locations. Sample sizes of Bluetooth data are not constant. Portable Bluetooth devices are susceptible to battery failure and data outages.

After reviewing many studies found in Table 1, several recommendations have been formed. 1) Bluetooth has proven to be an effective technology particularly for projects of a temporary nature. The ease of set up with limited cost also allows agencies to keep equipment on site and allows flexibility when one wants to extend a study's length. Installation of hardware is required for the collection of data, and data is only collected while the equipment is set up. 2) Bluetooth is effective for obtaining travel times. A unique MAC address needs to be detected by two or more receivers to obtain travel times. Bluetooth struggles at obtaining exact travel times, but the error is usually small relative to the distance traveled. For example, a trip of 2 minutes may be recorded as taking 2.1 minutes. Bluetooth has been used to differentiate modes but relies on travel times which have significant error in urban environments. As a result, Bluetooth is not the best technology for mode differentiation. 3) Given the nature of the data, true O-D networks are not feasible to implement, but cordons can be set up to detect travel patterns between entry and exit points and times. Increasing the area of the study will also increase bleed-off and reduce collection of useful data in and out of the cordon perimeter. Since true origin and destinations cannot be obtained, Bluetooth should not be used for long distance O-D studies. Overall, Bluetooth is a mature technology that can provide reliable data that can aid transportation agencies.

CHAPTER 3 - GLOBAL POSITIONING SYSTEMS

3.1 Overview

The Global Positioning System (GPS) is a network of satellites installed and operated by the United States Air Force which was made available for public use in the 1980s. By triangulating the signals reflected from multiple GPS satellites, GPS-enabled devices can accurately determine their position anywhere on earth independent of local infrastructure or data connectivity. This has made GPS an attractive technology for in-car position and navigation systems on consumer and commercial vehicles, and for managing fleet operations.

Commercial providers of GPS services – TomTom, HERE, INRIX, and others – can observe the locations and timestamps of devices on their networks by tracking GPS equipment places in vehicles and in some cases cellular apps on phones. They then sell this data in raw and aggregated forms for use in transportation studies and analysis. In some instances, cellular phone navigation apps use location-based services or other mobile-device locating features in addition to GPS information; the GPS data available to purchasers therefore could contain a mixture of GPS data and these additional technologies, which are discussed in more detail in the following chapter.

This chapter begins with an overview of GPS technology that informs a subsequent review of transportation studies conducted using GPS devices and cellular phone navigation apps. Lastly, a section of strengths and weaknesses will be summarized. The chapter ends with a discussion of scenarios in which it would be appropriate or not to use GPS data.

3.2 GPS Technology

GPS was originally developed as military technology used to identify precise targets and aid military vehicles in navigation. In the 1980s the United States military gave the public access to the system. Originally, civilian use of GPS was limited to a precision of 330 feet under a policy known as selective availability. A 2000 law eliminated selective availability, greatly enhancing the usefulness of GPS as a consumer navigation and wayfinding technology (U.S. DOT, 2014; Kamali, 2015; Huang, 2018; NOAA, 2018; GIS Geography, 2020). Today, with the right combination of weather, atmospheric conditions, and receiver technology, it is possible for the public to have precise location information.

GPS works by trilateration. In geometry, trilateration is the process of finding absolute or relative locations or points from a set of measured distances. For trilateration to occur, a minimum of three satellites is required to isolate a singular GPS device location. Since the position of all satellites can be determined at all times, a position on the surface of the earth relative to a group of satellites can be obtained. The distance between one satellite and a singular GPS device is easily calculated.

Figure 5 (a) and (b) show possible locations given by one satellite. With one satellite and a known distance between the satellite and a GPS device on earth's surface, an entire circumference of possible locations can be drawn on the surface of the earth.

Figure 5 (c) shows that with two satellites and known distances between the satellites and a GPS device on earth's surface, two overlapping circles can be drawn on the surface of the earth. The intersecting points of the two circles provide two possible locations where the GPS device is located on earth.

Figure 5 (d) shows that with three satellites and known distances between the satellites and a GPS device, three circles can be drawn on the surface of the earth. The singular point of overlap is where the GPS device is located. Each added satellite increases locational accuracy. The ability of GPS to continually track each GPS device makes it possible to identify exact traces of individuals. Satellites require a direct line of sight to calculate distances. Obstructions such as buildings or canyons can lead to reflection of GPS signals that can reduce positioning accuracy.

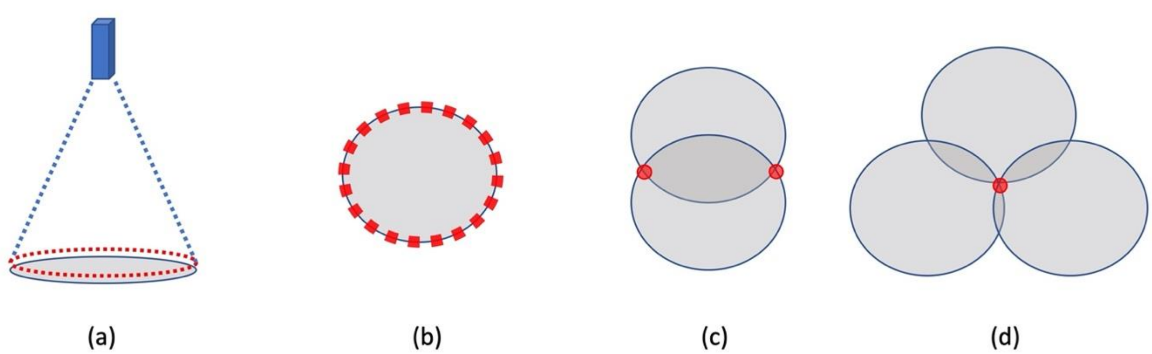

Figure 5: Diagrams depicting the geometry behind trilateration.

Due to the accuracy limitations of personal GPS-equipped devices, GPS traces often need to be realigned to match known road networks. GPS is highly accurate, but the error recorded from consumer devices can be up to 16 feet in open-sky conditions (Diggelen & Enge, 2015; Tomastik & Mokros, 2017). In situations where roads are parallel or near one another it becomes essential to match the GPS point to the correct road. Researchers such as Marinelli et al. (2017) have worked on methods to realign GPS points to known road networks. In this instance, the researchers were able to identify specific lanes a particular vehicle was traveling in near signalized intersections. The researchers were successful in identifying individual lanes and hope to use findings to assist adaptive signal changes based on real-time GPS data from mobile phones.

In literature, each recording within a trace has a location and associated time which is referred to as a "ping." In many cases, pings can be as frequent as once every second (Goodall, 2012; Tao, 2012), but such fine resolution may not always be necessary. Pinjari et al. (2014) reduced their data to 15-minute intervals; in this case, the data was over an entire state and commercial trucks typically travel on major freeways. Using a ping rate of 2 seconds would have created redundant data that would have been difficult to manually process. This study and similar ones like it have found that lower ping frequency is suitable for long distance studies. A much shorter ping rate would be needed for urban environments. Each GPS device recorded has a unique identifier. The unique identifier is assigned to the device at random and can be reassigned. Each unique identifier can only be assigned to one device at a time. Between identifier reassignments, researchers are able to trace the exact routes of a device with that unique identifier. For example, if a company reassigns unique identifiers every 48 hours to protect user privacy, then the useful data for any identifier must be collected within that 48-hour window before the next reset occurs. In literature, GPS data is often referred to as a probe or floating car data. Historically, single vehicles were used to identify travel times in a corridor, with the study vehicle referred to as a probe. As technology advanced and hundreds of vehicles could be tracked simultaneously, the term "probe" remained. This document will refer to any vendor-provided GPS data as either GPS probe data or simply GPS data.

Physical GPS devices are regularly installed in commercial fleet vehicles as commercial enterprises are most interested in evaluating fleet location and overall driver performance. Due to the large concentration of GPS equipment in fleet vehicles, some data vendors are able to sort commercial fleet vehicles from regular traffic. The heavy implementation of GPS in commercial fleet vehicles can help agencies to target specific modes for isolated studies; a good deal of research has been conducted on GPS data for taxi and long-haul trucking fleets, for example. On the other hand, the overrepresentation of commercial fleet vehicles in GPS may mean that general transportation findings are heavily biased towards these fleets.

Agencies commonly purchase GPS data for specific regions for specific times, e.g., for the state of Utah for the year of 2018. Often, an agency will need to negotiate what regions and data are included in a large purchase. For example, regions with higher populations produce more data and therefore vendors will often charge more for this data set. GPS data is sometimes sold unfiltered or in raw form to agencies. Other companies will sell data that is processed and will provide a subscription to a user interface tool. GPS data is packaged and sold in a large variety of formats and further inquiry should be done into prices for the state of Utah or smaller regions within the state.

Obtaining raw data from a vendor has many advantages but may overwhelm a transportation agency. Processing raw data is labor intensive and may be difficult for agencies to handle but could provide several advantages over other formats of GPS data. Raw data contains large amounts of extraneous information that must be extensively filtered out. It is recommended that DOTs partner with universities or third-party data experts when working with raw data formats. It may take several years to properly filter and analyze large data sets (A. Pinjari, personal communications, August 2020). Many data filtering techniques exist, though a specific description of each of these methods is outside the scope of this study. Kamali (2015) and Thakur (2015) have provided techniques to reduce the file size while maintaining data accuracy. Obtaining raw data has two main benefits: First, data can be custom filtered to meet department needs; and second, data may be layered with other data. Obtaining raw data may give agencies the ability to custom filter and manipulate data as circumstances change during a single project or use and revisit the original data in future projects. After custom filtering a data set, it is possible to overlap and integrate shape files, statistics, as well as other passive data sets. Many data vendors will not sell data in a raw format because of privacy concerns, or will sell data with the trip endpoints obscured (N. Markovic, personal communication).

One known form of composite GPS data is so-called *telematics* data, which combines GPS points with driver behavior data including acceleration and speed information. Telematics data is often collected by insurance companies and in some cases is sold to transportation agencies, though we found no evidence of specific studies using this data in North America. Lehmann et al. (2017) used a form of telematics to generate a model that estimated emissions more accurately due to the acceleration and deceleration patterns of vehicles. Bazzani et al. (2010) used the GPS portion of telematics data to identify traffic movements, speeds, and activity duration. The General Services Administration posted a document online showing the many ways government agencies have used telematics data to reduce budgets. Several agencies identified underutilized vehicles and reduced the size of their fleets (General Services Administration, 2014). Telematics data may be of interest for future research.

3.3 Transportation Studies using GPS Data

This section will explain how GPS data is being used in various transportation applications. The literature is sorted into four categories: network construction, travel times, O-D, and volume estimations. Each paragraph will explain a single application of GPS data and will also determine if GPS data is a good fit for that application.

3.3.1 Network Construction

The ability of GPS to continually track each GPS device makes it possible to identify exact traces of individuals. When GPS pings are aggregated and layered it is possible to infer road networks without an existing map. Figure 6 shows how agencies can infer road networks from unique GPS traces. Each color represents a unique GPS trace. Accuracy of each trace may be off slightly, but the aggregated results converge to identifiable paths. Image (a) shows what raw GPS points might look like before processing. Each dot contains longitudinal and latitudinal coordinates as well as an associated time. Image (b) shows possible roads placed where unique traces overlap, and (c) shows an inferred road network.

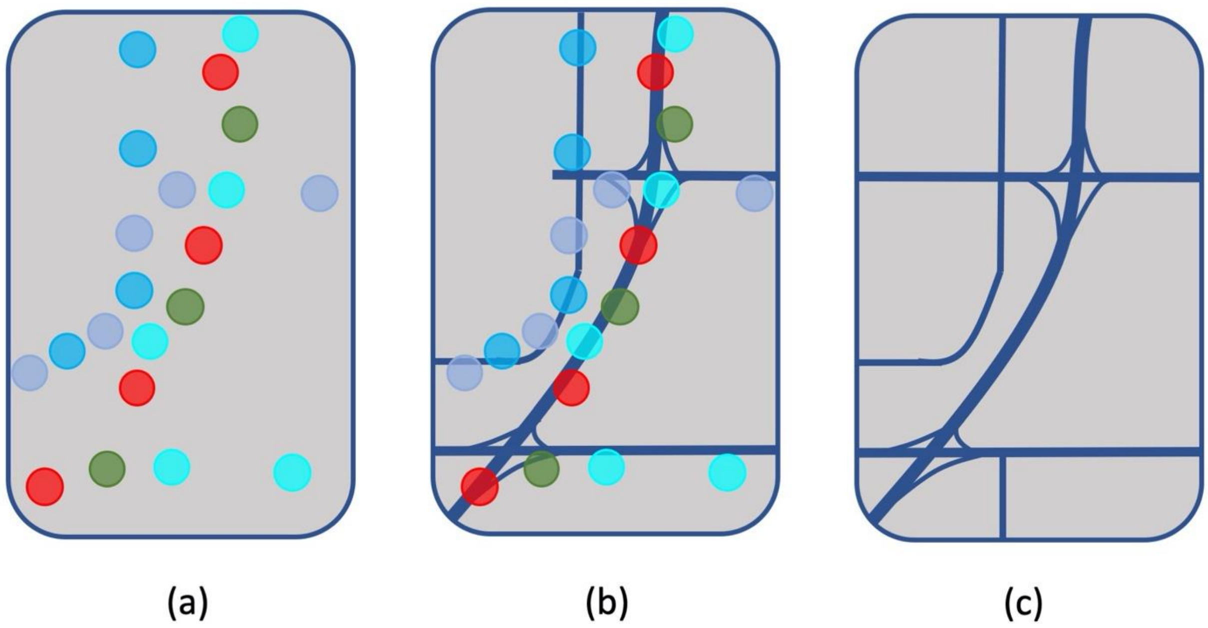

Figure 6. diagrams showing the evolution of GPS pings to constructed road networks.

Road networks are continually changing, making the task of creating up-to-date digital representations of highway assets extremely challenging. Whether it is a new development or a recent upgrade to an existing road, it is important to have up-to-date maps. Traditional methods for updating road networks, such as aerial photography and internal linear referencing data sets, are labor intensive and require a significant time commitment to implement. GPS data has the possibility of identifying new roadways and changes to existing roads almost instantly by observing the paths of vehicles who use the system. Elements of a map that can be obtained from GPS traces include the type of road, number of lanes, intersections, rough density information, direction of traffic, etc. HERE technologies, a GPS data vendor, is known to provide a service where they aggregate data from a fleet of LiDAR-equipped vehicles as well as GPS from navigation systems in personal vehicles (HERE Technologies, n.d.). Biagioni and Eriksson (2012) explained how various algorithms are used to identify road networks which are then validated using existing maps. In one algorithm, the data points were grouped into sections where the authors determined the center of the road by aggregating the traces of several different vehicles, which were then connected to identify a single road. Another method involved connecting GPS points with arcs and straight lines to calculate the centerlines of roadways. Once centerlines were calculated the lanes could be identified relative to the centerline. Davies et al. (2006) developed an algorithm that found road networks from aggregated GPS traces instead of treating each GPS trace individually. Clustering of aggregated traces allowed the authors to identify centerlines of roadways. The researchers discovered that this method created more accurate maps where larger GPS errors existed.

A number of researchers have also used GPS traces to identify intersection configuration. Xie et al. (2015) focused on identifying intersections with GPS traces. Each intersection was identified by tracking the change in direction of a small sample of 889 total GPS traces of a shuttle bus. This was done by detecting places in which multiple different directions could be taken. The authors indicated that a bend in a road has one possible change in direction where an intersection has at least two different directions. A ping rate of once every second allowed researchers to identify sudden changes in direction. The algorithm detected 36 of 33 intersections on the study route. The algorithm misidentified 3 locations as intersections but identified all 33 actual intersections. GPS noise, such as reflection from tall buildings, was attributed for the readings of three false intersections. Intersection accuracy was found to be off by an average of 74 feet. The researchers were able to detect direction of traffic and found that 83% of all roads in the study were correctly matched within 33 feet of the true location. These studies show that using GPS can identify road networks with reasonable accuracy but that further work may be needed to perfectly match GPS traces to actual road locations due to minor accuracy issues.

Further research has been done to automate network construction. Qiu and Wang (2016), Chen et al. (2016), Mariescu-Istodor and Franti (2018), and Bastani et al. (2019) have all worked to improve the accuracy of the aforementioned algorithms with minor improvements. Bastani et al. (2019) concluded that automatic network construction can reduce the intense labor required to maintain and update road networks. The researchers admitted that current network construction techniques have more errors than maps created by traditional map makers who use methods such as aerial photography or satellite imagery etc. Zhang et al. (2017) attempted to fully automate network construction but found that a human element is still required. The method fell short of identifying off ramps and other minor side paths, but it was able to identify the main roads and highways. Further research was needed to differentiate direction of traffic as well as obtain accurate lane information. Although GPS can and is being used to infer road networks, further research is needed to increase the accuracy of these automated methods to that of professional map makers who use traditional validation tools such as aerial photography or satellite imaging (Bastani, 2019).

3.3.2 Travel Times

Agencies can use GPS to determine both average and instantaneous speeds of motorized vehicles. Longitudinal and latitudinal coordinates of vehicles are recorded in regular intervals as pings. Travel times and average speeds between pings may then be derived, as the ping rate is relatively frequent. This section will describe how travel times have been used in various applications. Figure 7 depicts how a series of pings can determine average travel times between locations. Useful ping rates vary from several times per second to 15-minute intervals.

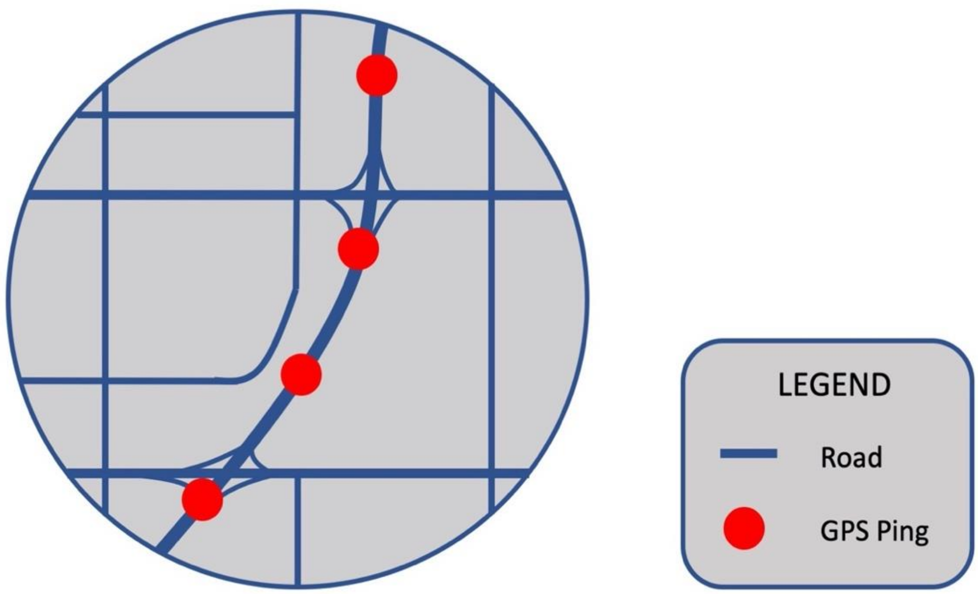

Figure 7. Travel times derived from GPS pings on a road network.

UDOT currently uses GPS data to derive travel times through iPeMS. UDOT purchased HERE data in 2017 and have used the iPeMS interface to determine travel times, delay estimations, holiday travel patterns, and more. The data has been used for prioritizing future projects, improving signal time performance, post-project analysis, and crash impacts. The interface has primarily been used to predict travel times and does not provide any volume estimation. The interface makes data more accessible to engineers not trained as data scientists (G. Farnsworth, Personal Communications, November 2019).

Real-time estimations of travel time can be obtained with GPS data. Tao et al. (2012) used real-time mobile phone GPS data as an input into a simulation to find microscopic speeds from corridor average travel times. Data was filtered and then assigned to the closest node within a small study region of Copenhagen, Denmark to find link speeds. By aggregating the average speeds of vehicles, the authors were able to determine average travel speeds for every 10-minute period on each link of the entire Copenhagen road network. This real-time application could be used as an analysis tool to monitor the flow of traffic.

GPS has the ability to identify if an individual is speeding. GPS can track average speeds between points and in many cases spot speeds at an instantaneous moment. Markovic et al. (2018) investigated how GPS data could help DOTs identify where drivers commonly speed. By analyzing GPS data either spot or average speeds could be determined and then used as a comparison to posted speed limits for that road segment. From this information a variety of interventions can be taken to improve public safety by encouraging drivers to reduce speeds. Understanding roadway use could help with assessing maintenance and roadway improvements that encourage lower speeds such as speed bumps, radar speed signs, law enforcement etc.

Researchers are investigating how GPS can identify accidents or sudden events on highways. Stimpancic et al. (2016) created a congestion index and suggested that areas with high congestion variability could indicate where accidents or construction are occurring. Asakura et al. (2015) tried to identify incidents on freeways using GPS data. Two algorithms were designed to detect sudden incidents. The first algorithm used only individual travel times. The second algorithm relied on a minimum of three consecutive vehicles to detect the shockwave of a sudden incident. In theory, one vehicle at free flow would precede the accident and the following vehicles would encounter reduced speeds and therefore a shockwave could be calculated from the rates at which the following vehicles slowed down relative to the first vehicle. They found that GPS could present false readings and miss incidents that occur if no congestion or delay was detected. This paper was able to identify traffic incidents, but future work is required to refine the methods and eliminate false alarms.

GPS has been used to monitor and classify congestion. In Ankara, Turkey, Altintasi et al. (2016) used GPS data to create a model that could determine travel-time estimations for every minute interval. The researchers used travel times to detect where congestion or delay was accumulating in real time. Altintasi et al. (2016) compiled average travel times for each one-minute interval along a busy corridor and translated average speeds into level of service (LOS). Once LOS A through F was assigned, they could determine where the majority of congestion occurred. Using only travel time as a parameter, they determined that GPS data is able to detect intersection congestion, bus stops, and other major points along a corridor. The researchers could also detect when bottlenecks started, ended, and identify the length of queue within a bottleneck. This paper provided evidence that having a frequent ping rate – 1-minute intervals – provides sufficient detail for identifying traffic patterns in real time. Altintasi et al. (2016) found low GPS penetration rates of around 3% were sufficient for their purposes which undercuts the minimum 5-7% penetration rates suggested by Tao et al. (2012). In Quebec City, Canada, Stimpancic et al. (2016) used GPS data collected from a mobile application to track real-time traffic patterns. Their 6% penetration rate failed to include many arterial and residential streets. Despite the lack of spatial resolution, the researchers were able to create a congestion index number that provided some level of accuracy to true congestion levels. Congestion index was calculated as $\frac{(Freeflow\ Speed - Actual\ Speed)}{Freeflow\ Speed}$. This created a scale where a value of 1 would indicate complete congestion and a value of 0 would indicate free flow. Researchers confirmed that congestion exists in specific areas at specific times and varies throughout the day. Data collected from the application could highlight patterns such as the ripple of morning suburban traffic approaching the downtown region. They found that GPS data could be used as a real-time analysis tool to identify congestion index numbers of all city roads simultaneously. In addition, one could infer that consistent congestion could highlight areas that need upgrades while areas with high congestion variability could suggest where accidents or construction might be occurring. The researchers indicated that they lacked sufficient sample sizes on many arterial and residential streets. Further research is necessary to fill spatial gaps on arterial roads and then refine methods.

Travel-time reliability is an important performance measure used by logistics companies. The U.S. economy relies heavily on the transportation of goods of which roughly 70% of tonnage travels by truck (Kamali, 2016). Businesses often prefer accurate travel-time forecasts in contrast to the chance that they will have a quick delivery route. To ensure smooth product delivery, companies require accurate trip time estimations for any given hour of the day. As noted earlier, Stimpancic et al. (2016) was able to create a congestion index from 6% penetration rate of total traffic. This index was able to show real-time congestion for the entire Quebec City road network. Pinjari (2014) used a large commercial truck GPS data set to identify the average travel time for each mile of Florida highway. Extensive processing and work with a geographic information system (GIS) allowed the researcher to create a bidirectional map displaying average speeds for five time periods in a single day. This work was then expanded to create a travel-time index for planning efficient routes through Florida. Golias et al. (2012) used a combination of GIS shapefiles and GPS to identify estimated travel times for each hour of the day. The researcher used GIS files to divide a highway into one-mile sections and then found the average travel time for each hour. These methods provide travel-time reliability that can be used by businesses to make shipments efficiently.

Bottlenecks can be determined and ranked by identifying travel times. McCormack et al. (2011) used travel-time information to classify bottlenecks in Washington state. Travel-time reliability was needed to accurately rank the worst bottlenecks. They successfully identified when a truck’s speed fell below 60% of the speed limit posted. This threshold suggested congestion in the area of study. The researchers started by creating a scale with three categories: reliably slow, reliably fast, and unreliable. The researchers calculated a travel reliability index by the formula $\frac{\#\ of\ periods\ where\ freeway\ is\ unreliable\ or\ unreliably\ slow}{total\ \#\ of\ periods}$. In addition to travel-time predictions, the authors were able to provide a detailed report which included location, length of segment, daily truck volumes, average speed, and travel-time reliability. Liao (2014) used a similar threshold value – usually 45 mph – to rank bottlenecks in the Twin Cities, Minnesota region. The congestion index was determined by the number of hours below the set threshold value during peak periods. Both studies were successful in identifying top bottlenecks in their regions and finding travel times or travel-time reliability estimations for each hour of the day.

GPS data can identify the duration that a vehicle remains stationary. Some congested ports and intermodal hubs impose fines for long turnaround times. These facilities intend on limiting truck queues and congestion. Golias et al. (2012) used GPS data to successfully track turnaround times of trucks within train-truck intermodal hubs. The use of GPS helped increase efficiency and reduce congestion inside intermodal facilities.

Supplemental GPS data is often used in simulations to determine how to implement traffic signal changes. Although simulations fail to account for all possible variables, insight into traffic flow at intersections and highways can be obtained. In New York State, Ban et al. (2011) explored how travel time, in a left-turning scenario, could help to determine queue lengths. Collected data could be used to influence left-turn signal timing. The speed of vehicles in a left-turning lane was used to identify queue lengths at a signalized intersection. The goal for the researcher was to have a consistent flow of traffic referred to as queue rear no-delay arrival time. This real-time measurement provides an alternative to traffic turn signal timing since traffic can be monitored before entering the vicinity of an intersection. This proactive method could better coordinate signal timing at intersections and reduce waiting times for all travelers. In Calgary, Alberta, Canada, Kattan et al. (2012) used software simulations to model effects of implementing GPS data for ramp metering. The researcher used three distinct ramp metering test scenarios: pretimed intervals, physical detectors, and probe-based data. Travel times on the main highway were determined exclusively using GPS data. GPS data was then combined with information collected from two physical detectors on each on-ramp to determine the rate at which traffic could be admitted onto the freeway. The researcher found that using probe data as a supplement outperformed exclusively pretimed or physical detector methods. With only 3% penetration rate as compared to pretimed or detector-based methods, traffic delay was reduced 7%, travel time was reduced 2%, and average speed was increased 1%. The researcher noted that increased GPS penetration rates had increased effectiveness at reducing delay and travel times and that penetration rates under 3% began to be less effective than pretimed or detector-based methods alone. Future work to test ramp metering in real-life scenarios has yet to be done.

3.3.3 Origin-Destination Analysis

Several studies have conducted research to prove the effectiveness of using GPS to find true O-D information. GPS traces have the ability to identify the actual start and end points in an individual trace. Figure 8 represents a typical GPS trace. GPS allows agencies to identify trip purpose because home and work locations can be inferred through repeated measurements.

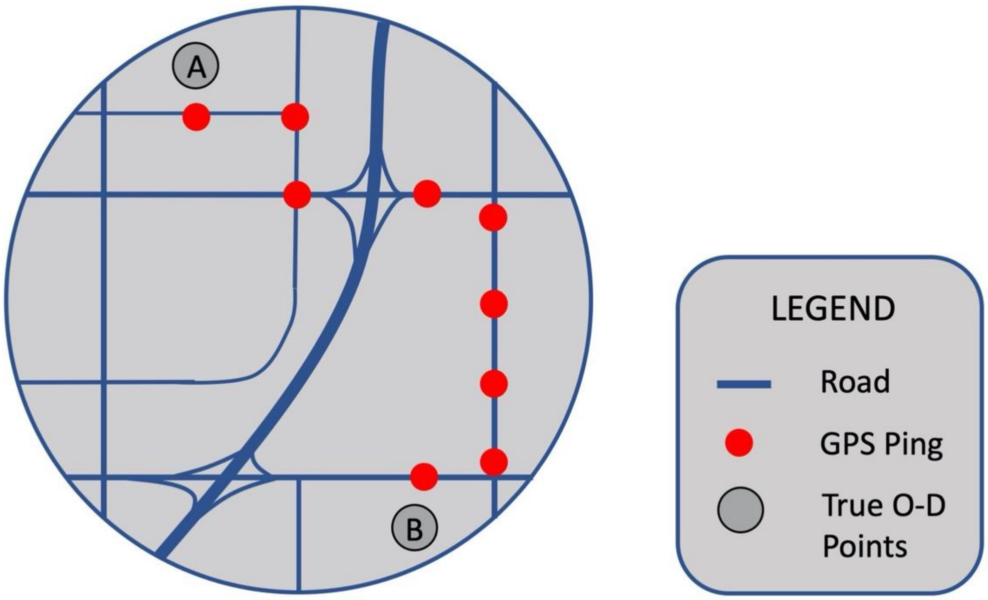

Figure 8. A series of GPS pings revealing a specific trace

Using GPS data, agencies can determine common routes traveled by drivers. Kamali et al. (2015) worked to find the full diversity of route choice from O-D records. The researchers used an algorithm to identify trips with less than 75% overlap and then classified them as unique routes. In another case study, Kamali et al. (2015) found methods to isolate tanker trucks from a sample of commercial truck GPS data. GIS software was used to locate and identify petrol sites such as gas stations and delivery ports. Based on proximity of individual trucks within a geofence around fueling sites, they successfully identified tanker trucks from the entire truck sample. After isolating tanker trucks from the overall truck sample, the common routes of tanker trucks were identified. This method required significant processing and cooperation with the data vendor. Applying these techniques in other regions would require similar customized efforts. This type of study could allow one to identify which highways need future repairs amongst other applications such as targeted taxes or tolls. In related work, Tahlyan et al. (2017) used GPS O-D data to find the most relevant unique trucking routes. They discovered that short-haul trucks generally have more diverse route choice than long-haul trucks. They were successful in identifying most relevant and popular trucking routes.

Analyzing individual vehicle traces could reveal trip purpose. Markovic et al. (2018) was able to infer that a certain percentage of trucks had taken possible detours in the vicinity of weigh-in-motion (WIM) stations. WIM stations track commercial vehicles to ensure maximum weight-per-axle requirements are being followed. It was observed that increased vehicle miles were driven in indirect paths near WIM stations. Although the percentage of traffic suspected of diverting was only around 1%, the detection of truckers possibly avoiding WIM stations could identify unanticipated damages and environmental impacts to side roads from increased vehicle miles traveled (VMT). Knowing where commercial trucks are traveling off of main route highways could help DOTs enforce weight restrictions.

GPS data has been used for travel demand modeling in place of traditional travel surveys. Bernardin et al. (2014) used truck GPS data to develop a statewide truck traffic demand model. Since the data set represents a specific sample, it comes with an inherent bias. They found that short-haul trucks were underrepresented, so they developed expansion factors that could account for this sample bias. Markovic et al. (2018) found that O-D data was able to identify the number of trips in and out of the state of Maryland as well as county-to-county percentages. In this study, the sample included all vehicles with commercial trucks included. Markovic et al. (2018) also investigated how O-D information collected from 20 million traces could be used to identify potential transit deficiencies. The researchers took the 20 million data points and layered them over existing transit maps. The layered map revealed several routes that were frequently traveled and did not have existing transit service. This suggests that GPS may be used as an analysis tool to determine how to efficiently allocate resources as well as target future projects. In Florence, Italy, Bazzani et al. (2010) obtained telematics data that could identify instantaneous speeds, trip lengths, and duration of stays. Although the researcher was not trying to solve transportation problems, they found that GPS data could show individual movements within an entire city. From following the individual traces of vehicles, the researchers could infer home and work locations. In one analysis they created a roadmap showing three categories of average speeds. Having access to instantaneous speeds could reveal main thoroughfares or areas with congestion based on LOS definitions. In another analysis the researchers used statistical models to estimate the average downtime spent at daily activities. These types of analyses could help planners improve infrastructure such as parking facilities or transit.

It has been suggested that GPS O-D data could assist governments in identifying statewide commodity flow. Florida is known for being a consumption state with a higher import-to-export ratio. The imbalance of imported goods adds vehicle miles traveled (VMT) and as a result increases consumer costs. Zhao et al. (2020) investigated commodity flow by identifying empty trucks traveling through the state. Using Florida’s WIM stations, the researchers created a travel model that inferred empty trucks based on weight. Using data from WIM the researchers could identify rough percentages and headings of vehicles from WIM stations and then used GPS data to identify commercial trucking patterns in the state. The combination of these two data sources allowed the researchers to estimate where imports were coming from as well as where empty trucks were heading. This data could be used to inform governments to invest in industries that could help balance commodity flows in areas with one-way shipments. The researchers indicated that they were able to predict such flow, but that further research would improve the model.

Aggregated GPS data can help pedestrians find the best evacuation route during natural disasters. Ikeda et al. (2016) used real-time mobile GPS traces to identify safe pedestrian passage through an area. The most efficient path was generated after data was aggregated in real time and sent back to an application that the researchers were working with. The most efficient path was not always the shortest path. One example is where stairs or damaged roads were detected. This could have benefits for those in wheelchairs or those who have other conditions that affect which route to take. This technology could also help in a multitude of other natural disaster situations such as earthquakes or floods and does not seem to be limited to pedestrians.

Several researchers have tried to use taxi GPS data to estimate vehicle emissions. Smog and growing awareness of the negative effects of air pollution have sparked interest in identifying and quantifying air pollutants. Many researchers including, Liu et al. (2013), Weng et al. (2017), Zhang et al. (2017), Gately et al. (2017), and Kan et al. (2018) have investigated how to use GPS data as a backdrop for estimating emissions. Finding accurate emission estimates is difficult because of the multiple factors that affect the quantity of emissions produced. Researchers also need to know how large their sample is and how to expand that to the total population. The O-D shows the distance and time traveled, and the sample size is used to represent driving behavior of the entire population. It is important to understand that different vehicle activities will produce higher quantities of particular pollutants. Liu et al. (2013) explains that idling vehicles produce more PM_{25}, CO_2, and HC emissions while accelerating vehicles will produce predominantly CO emissions. This paper indicates the importance to not only understand O-D information, but driver behavior of a given trace. Driver behavior is largely unpredictable and can introduce significant errors or bias. Weng et al. (2017) indicated that speed played a part in the amount of emissions and what emissions were produced. This paper divided each trace into subsections that represented acceleration, deceleration, or cruise. Each classification was weighted differently. In Massachusetts, Gately et al. (2017) used travel times to identify how much emissions were produced by vehicles at different speeds. They then calculated the impacts of engine starts and transposed findings onto estimated fuel consumptions from various travel speeds. The researchers concluded that congestion had a modest impact on overall emissions. Vehicle models, regional standards, weight, and age could also alter emission results. Liu et al. (2013) indicated that emissions studies are often limited to local areas where standards and conditions are the same. If acceptable methods are achieved; transferring methodologies to other regions is extremely difficult. Liu et al. (2013) and Kan et al. (2018) researched how driving behavior could affect accuracy of emissions results. Kan et al. (2018) used methods to identify restarts, idling, acceleration, etc., to better represent true emissions from individual vehicles. Ultimately the error and lack of confidence in all of the studies leaves considerable room for future research. Telematics data, which is often collected by insurance companies to track driver behavior, was briefly discussed. It may be of worth to investigate how telematics data may supplement the missing driver behavior information required in these studies.

3.3.4 Volume Estimations

Traffic counts are an essential performance measure collected by agencies. Many researchers are investigating if GPS or other passive data forms could replace traditional traffic counts. Figure 9 depicts a simplified scenario where a physical counter detects 30 unique vehicles within a specific time frame. This information is used as a reference for GPS data that is collected at the same time and place as the physical counter data. In this example there is a ratio of 3 GPS traces to 30 actual vehicles (a penetration rate of 10%). In order to use the GPS points, the data must be scaled to represent the entire population. In this case a factor of 10 would be used to boost the sample to represent the entire population.

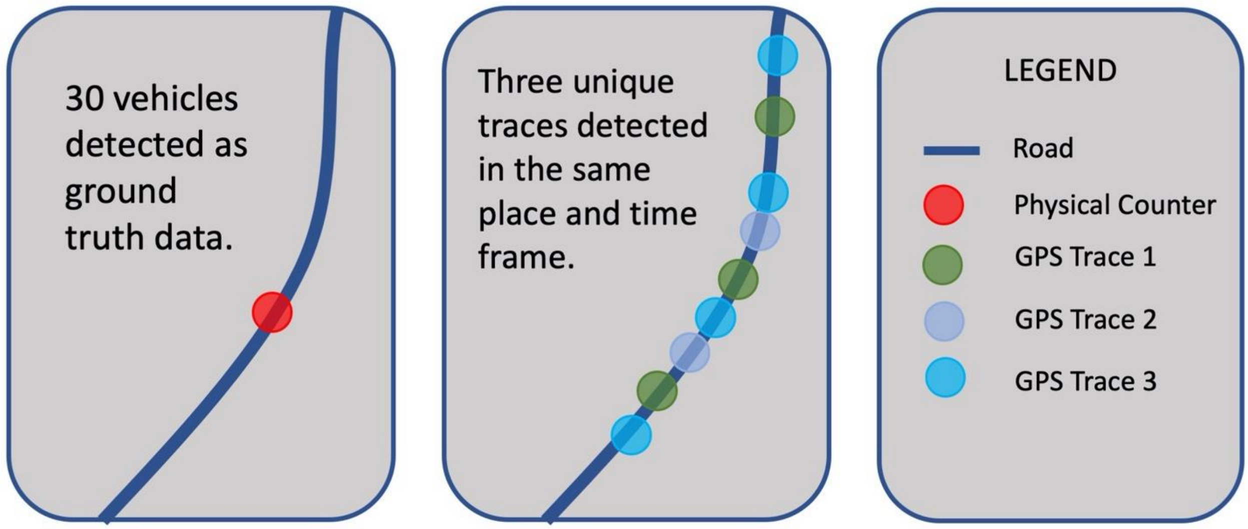

Figure 9. Example of how GPS traffic counts are validated and expanded.

Counting traffic is an important performance measure which is required in the United States. Traditionally, manual counts, pneumatic tubes, and in-pavement detectors are used to obtain traffic counts. GPS data currently has failed to provide reliable traffic count information as it represents a sample of all vehicles which is too small. Young et al. (2017) indicated that probe data is currently not robust enough to provide reliable traffic counts. Young et al. (2017) used commercially purchased GPS data to supplement existing methods. The authors found that GPS data combined with other sources such as loop detectors, counts, etc., had the ability to create results that came closer to true annual average daily traffic (AADT). It was also noted that error dropped significantly with the addition of GPS data. Using machine learning, Sekula et al. (2018) also observed that GPS data was an important input for their model because it significantly increased the accuracy of their model. Despite accuracy improvements, significant accuracy errors still existed and needed additional refinement. It should be noted that the researchers were able to obtain results that aligned closely with Young et al. (2017), but with only 1.8% penetration rates. Continuing the work of Sekula et al. (2018), Markovic et al. (2018) used machine learning in conjunction with GPS data to more accurately determine traffic counts but found that additional work was needed to accept results as a replacement to traffic counts. The researchers suggest mixing MDD and GPS to obtain better results. Another study, Miller et al. (2017) improved on various existing travel models and obtained a higher penetration rate of approximately 2.8%. They discovered that error from previous methods was reduced up to 45% and that the median error relative to ground truth data was approximately 18%. Miller, et al. (2017) indicated that ground truth data was collected from 296 automatic traffic recorder stations scattered throughout Utah. The researchers suggest that the new methodology could reduce the need for short-term traffic counts. Another researcher indicated that GPS data is unlikely to replace traffic counts completely, but that passive data has the potential of significantly reducing the regularity or scale of such traditional counts. Passive data sources combined with traditional counting methods provide DOTs with a closer representation of the truth (V. Bernardin, personal communication, July 2020).

Car-following models have been used to identify non-GPS-equipped vehicles. Goodall et al. (2012) used the characteristics of equipped GPS probe vehicles to detect the quantity, coordinates, and speeds of non-GPS-equipped vehicles. The researchers used prior car-following models that relate the behavior of how upstream traffic reacts to vehicles in front of them to locate and quantify unequipped vehicles. The researchers used known GPS data as an input to identify lane density and congestion. Their methodology failed to work in free-flow conditions because it required delay or congestion to identify vehicles without GPS equipment. This could be an alternative method of increasing sample rates and understanding traffic patterns on congested or arterial roads.

GPS data is used extensively to analyze commercial truck parking demand. Commercial trucking is the predominant means of delivering freight in the United States. In the United States there are hours-of-service regulations that prohibit truck drivers from driving over 14 hours (Diaz-Corro, 2019). As a result, truckers are required by law to take a 10-hour break between driving shifts. For long-haul trucks this means that drivers need legal parking for extended periods of time. In many places truck parking is in high demand and often there is inadequate space for them to park legally. Trucks may be found on freeway exit ramps or parked on the side of the highway which poses significant safety risk to all parties. Arkansas has an annual program where they have conducted regular overnight truck-stop demand analysis via manual counts. Diaz-Corro et al. (2019) used this observational data as comparison to create expansion factors on GPS data to create a parking demand model. Given the extensive log of previous parking count studies, the researchers were able to make a model that resembled data collected in the annual manual count studies. Many states may not have prior data records which would make this method difficult to validate. One caveat is that since this study used samples that typically did not fall on season peak periods, it failed to account for variability in seasonal demand. In another study, Torrey (2017) used GPS data to locate trends for truck parking demand and determined how various truck stops were over capacity. Oregon DOT (2019) is looking into clustering of trucks in non-designated parking as well as government-provided roadside facilities. They hope to use results as performance measures to isolate future project prioritization. They also hope to develop a demand forecasting model that will assist them in planning into 2040.

Sometimes there is adequate truck parking, but drivers may pass by because of vacancy uncertainty. Haque et al. (2016) used GPS to identify truck driver behavior. They found that number of lanes on the freeway, number of trucks parked on the on-/off-ramps, among other metrics affected whether truckers upstream would stop at a rest area or continue to drive on. They hope to use these results in designing better functioning rest stop facilities in the future that will attract truckers and clearly identify parking vacancies.

3.4 Analysis of GPS as a Data Source

This section will discuss the strengths and weaknesses of GPS technology so that agencies will understand how to best implement the technology. The section will first discuss trends in penetration rate and then will summarize strengths and weaknesses.

3.4.1 Penetration Rates

Obtaining sufficient GPS penetration rates is critical for obtaining accurate results. Generally, coverage is not an issue, but obtaining a penetration rate high enough is. Penetration rates are particularly low in less traveled areas such as arterials or rural locations (Stimpancic, 2016). It is noted that penetration rates for various vendors have increased over the years (N. Markovic, personal communication, July 2019). The majority of algorithms require a minimum penetration rate to provide reliable analysis. Tao (2012) found that 5% penetration rate on highways and 7% on arterials was sufficient for analysis. Without the required penetration rates, GPS data has a limited use which makes the data less valuable to agencies. In contrast to Tao et al. (2012), Markovic et al. (2018) found that penetration rates could be as low as only half a percent and as high as 5.5% when they combined machine learning techniques for O-D applications. It appears that minimum penetration rates required may depend on the travel model being used. Agencies should asses if the penetration rates available are sufficient for department purposes.

It is important to understand how filtering a raw data set may reduce the sample size and thus affect the final penetration rate of a sample. For instance, sometimes GPS data has large temporal or spatial gaps which may need to be removed before its use. Tahlyan et al. (2017) found that 50% of the derived trips from raw data had spatial gaps that were large enough to miss network links therefore making analysis difficult. In instances such as this, 50% of obtained data may need to be filtered out before its use. The amount of filtering will depend on the application of the agency.

Using lower penetration rates could increase the chances of magnifying a bias. Each sample is expanded to represent the overall population. If a characteristic in a sample exists, which is not uniform throughout the total population, expansion of that sample leads to overrepresentation of that characteristic. The overrepresentation of any characteristic in a sample is known as a bias. For this reason, some applications require a higher sample rate for better accuracy. Historically, travel times require a smaller sample to accurately represent an entire population, but O-D applications often require larger sample rates to be considered valid for representing entire populations.

Table 2 explores penetration rates of the major studies explored in this literature review. The chart below is ordered chronologically by general applications. GPS penetration rates do not appear to be dependent on any particular factor other than the vendor who supplied the data. Vendor information is not included in Table 2, but the sample source column may help to identify different GPS data sets.

Table 2 Penetration Rates Observed in GPS Studies

Author	Year	Penetration Rate	Sample Source	Success	Notes
Network Construction					
Davies et al.	2006	-	-	Yes	Network Construction
Biaglioni & Eriksson	2012	-	-	Yes	Network Construction
Xie et al.	2015	-	Shuttle Buses	Yes	Network Construction
Zhang et al.	2017	-	Taxis	Partial	Network Construction
Marinelli et al.	2017	-	All Vehicles	Yes	Map Matching
Bastani et al.	2019	-	-	Yes	Network Construction
Travel Times					
McCormack et al.	2011	-	Trucks	Yes	Travel-Time Reliability
Ban et al.	2011	-	All Vehicles	Inconclusive	Traffic Signal Queue Length
Tao et al	2012	-	All Vehicles	Yes	Speeds
Golias et al.	2012	-	Trucks	Yes	Truck Turn Time
Kattan et al.	2012	3%	Cars	Yes	Ramp Metering
Liao	2014	10.0%	Trucks	Yes	Travel-Time Reliability
Pinjari et al.	2014	10.1%	Trucks	Yes	Travel-Time Reliability
Asakura et al.	2015	0.2%-0.5%	All Vehicles	No	Real-Time Congestion
Altintasi et al.	2016	3%	All Vehicles	Yes	Real-Time Travel Times
Stimpancic et al.	2016	6%	All Vehicles	Partial	Real-Time Congestion
Origin Destination					
Bazzani et al.	2010	2%	All Vehicles	Yes	Route Choice - Telematics
Liu et al.	2013	4.3 - 8.1%	Taxis	No	Vehicle Emissions
Bernardin et al.	2014	10.1%	Trucks	Yes	Truck Demand Model
Kamali et al.	2015	10.1%	Trucks	Yes	Commodity Isolation
Kamali et al.	2015	10.1%	Trucks	Yes	Route Choice
Ikeda et al.	2016	-	Pedestrians	Yes	Evacuations Routes
Tahlyan et al.	2017	-	Trucks	Yes	Route Choice
Weng et al.	2017	-	Taxis	Partial	Vehicle Emissions
Zhang et al.	2017	-	Taxis	No	Vehicle Emissions
Gately et al.	2017	-	All Vehicles	Inconclusive	Vehicle Emissions
Kan et al.	2018	44%	Taxis	No	Vehicle Emissions
Markovic et al.	2018	1.8%	All Vehicles	Yes	Transit Improvements
Markovic et al.	2018	1.8%	All Vehicles	Yes	Various Applications
Zhao, et al.	2020	-	Trucks	Partial	Commodity Flow Analysis
Volume Estimations					
Goodall et al.	2012	-	All Vehicles	No	Vehicle Detection
Haque et al.	2016	3-20%	Trucks	Yes	Truck Parking Demand
Torrey	2017	-	Trucks	Yes	Truck Parking Demand
Young et al.	2017	0.57%	All Vehicles	No	Traffic Counts
Miller, et al.	2017	2.8%	All Vehicles	Partial	Traffic Counts
Sekula et al.	2018	1.8%	All Vehicles	No	Traffic Counts
Markovic et al.	2018	1.8%	All Vehicles	No	Traffic Counts
Diaz-Corro et al.	2019	-	Trucks	Partial	Truck Parking Demand

3.4.2 Strengths

Satellites orbit the earth uninterrupted which enables companies to track locations of individuals continuously. GPS data is often available for retroactive purchase in the event that an agency needs comparison data sets. When GPS data is collected by companies it is not restricted to one small region as is Bluetooth data. The data represents a sample of the entire available population at every possible location where GPS connections exist.

Agencies do not need to install, maintain, or set up equipment for GPS data collection. GPS vendors collect data on behalf of agencies. GPS infrastructure exists because of the efforts of the United States military and the collective purchase of smartphones with GPS functionality. The infrastructure is highly reliable and properly maintained without any involvement from transportation agencies.

GPS is highly accurate and can track entire individual traces. The average GPS-enabled smartphone can introduce an accuracy error of 12 to 16 feet in open sky conditions (Diggelen & Enge, 2015; Tomastik & Mokros, 2017). Accurate location information paired with frequent ping rates allows individual traces to be identified. Unlike Bluetooth which tracks vehicles as they pass a receiver, GPS is being tracked continually at regular intervals. Pings can be as frequent as once every second (Goodall, 2012; Tao, 2012; Asakura, 2015). One researcher used rates with up to 15-minute intervals (Pinjari, 2014). Chapter 4 will explain that LBS can also have similar rates but is known to be more sporadic. GPS tends to have pings with consistent intervals that provide accurate locations. The regular pings and accuracy of GPS locating allows for true O-D to be obtained. Each ping has an associated location that can reveal entire traces that include the true start and end of each trip. A complete trace has the potential to help agencies infer trip purpose because home and work locations can be inferred.

Instantaneous speed information is frequently collected for each ping. From the location and associated times, average speeds may be obtained between two consecutive points. This allows additional flexibility for agencies who prefer one over the other.

Several vendors install GPS equipment in fleet vehicles which makes it easier for agencies to obtain data sets specific to a particular mode. In addition, some data vendors are able to separate fleet vehicle data from regular drivers because of how data is obtained by these vendors. In general, GPS data can identify motor vehicles well but data for active modes seems to be in short supply. Additional research should be conducted to see how GPS could be used for transit and active transportation modes.

3.4.3 Weaknesses

Penetration rates for GPS data are generally lower than Bluetooth or MDD. Table 2 shows that penetration rates for an entire population (All Vehicles) are often around 1%-2%. The low penetration rate may make GPS data unusable for certain applications such as extracting O-D. If one wants to isolate a particular mode such as trucks or taxis, Table 2 shows the penetration rate of a subsample for trucks and taxis. One data set consistently sampled 10% of all trucks in North America, while taxi samples range from 4% - 40% of all taxis in a specific region.

Finding the right ping rate is important for many projects and sometimes the ping rate may not be frequent enough for use. GPS technology has the ability to record longitudinal and latitudinal coordinates on an interval basis. Some data sets record locations as often as every second (Tao, 2012). If timestamps do not occur frequently enough then it is difficult to determine which route a vehicle traveled as there could be several alternative routes between recorded times. Kamali et al. (2015) used data with 5-20-minute intervals between timestamps and found that in many cases they could not accurately determine speeds or even route traveled. For example, a 15-minute ping rate can increase the number of possible parallel routes traveled during that time while a 1-second ping rate would leave no room for guessing. Tao et al. (2012) discovered that 10 seconds between timestamps was optimal for eliminating alternative routes and avoiding excess data collection. Overly frequent timestamps created data clutter that wasted server storage yet resembled data with smaller ping rates. In some instances, there is insufficient data collected for arterials and predictions cannot be made. Often areas that have insufficient data are of lower priority or concern. It may be better to have higher ping rates than needed because one can always filter out excess information.

GPS can lose signal underground, near tall buildings, or near natural features such as canyons. GPS requires direct line of sight between satellites and the GPS device. Reflection can cause the signal to bounce which causes issues with accurately identifying where a GPS device is truly located. Additionally, signal might be lost when vehicles are in tunnels or structures such as parking garages. A loss in signal can result in one trip being split into two which generates a false number of total trips in a study area. Pinjari et al. (2014) worked on a method that combined pieces of trips spliced by a loss of signal. This method was successful in matching commercial trucking trips that were incorrectly spliced.

In the past, GPS has not been accurate in urban environments. Prior to 2014, several major probe data vendors provided accuracy on highways but struggled to provide reliable data in areas with traffic signals. Changes to algorithms have drastically changed the accuracy of results in just a few years (Sharifi, 2017). Findings indicate that probe data is now sufficient to predict arterial patterns with reliable accuracy.

Raw GPS data requires extensive filtering which may be out of the scope of a transportation agency. Kamali et al. (2015) described the monstrous task of filtering 725,000 commercial truck GPS data points. After significant processing, the original data set was shrunk to approximately 84,000 trips. Information that needed to be removed included outliers as well as extraneous information. An example from this paper explained that when a truck stops for refueling or other activities a single trip may be divided into two separate trips. The unnecessary divisions falsely represent the true origin and destination. Another example from the paper mentioned cyclical delivery routes that start and end at the same place. In this scenario, it is best to divide each delivery as a separate trip. Sometimes anomalies must be removed. An example of an anomaly in the data is having a truck speed above 100 miles an hour. In the United States trucks are often limited to 60 miles an hour. Trucks traveling at unreasonably high speeds need to be removed from the sample. After sifting data, the remaining points needed to be map-matched onto a known road network. This level of processing may require third-party help depending on an agency’s staffing and resources. This level of processing could take several years to develop a system, train staff, process data, and perform meaningful analysis. The computing power needed to store and process that much information could be a limiting factor for many smaller agencies. A partnership with universities or a third-party data specialist may be a better option for agencies.

When purchasing filtered data engineers should know how it was filtered. Pinjari et al. (2014) explains that knowing what data you are given is essential. For instance, the travel-time data that you purchase might include refueling and overnight rest stops. If you want to know total driving time excluding refueling and other stops, then the data needs to be filtered differently. Knowing how the company collects their information can be important based on what you plan to use the data for. It was also mentioned that some GPS devices are programmed to stop data collection if the vehicle is turned off or does not change position for a certain length of time. These variations could affect the precision of the data and should be known to the purchaser of the data sets.

GPS data is more prone to show bias. It is important to remember that GPS data sets represent a sample. Lower penetration rates may increase the chances of unintentionally introducing bias to a scaled-up sample. For example, GPS devices are installed in fleet vehicles on a mass scale. Bernardin et al. (2014) found that the probe data used was skewed towards long-haul trucks and that short-haul trucks were underrepresented because of how data from fleet vehicles were collected. This is particularly problematic because driving habits differ between commercial and consumer vehicle owners. In addition, a consumer is unlikely to use GPS to commute to work or travel to frequented places. Instead, the consumer will use GPS for irregular trips which will introduce a bias to irregular locations. Transportation analysts want data that represents the travel patterns of individuals throughout their daily routine and not just the irregular trips. When samples that contain bias are scaled up, the bias becomes magnified. Bias could also hyper represent certain professions, socioeconomic groups, and age groups. Inversely, bias could exclude various minorities such as elderly individuals without smartphones. To avoid such problems larger samples are ideal especially for applications that involve O-D matrices. Other biases may exist, and further research should be conducted to identify all known biases.

The accurate nature of GPS, and its ability to track true O-D, open the door for potential privacy issues. GPS is accurate and can track true O-D information which can reveal homes and workplaces. Companies that sell data often try to obfuscate true O-D information by simplifying the end points to traffic analysis zones (TAZs) or counties. This makes the data more anonymous but it is not fully secure. By identifying a unique workplace and home location, inferences can be made to identify an individual. If one acquired multiple data sets and superimposed the data, one might be able to identify individuals based on predictable travel patterns (Thompson, 2019).

3.5 Summary and Recommendations

GPS is a well-established technology that relies on a network of satellites orbiting the earth. Signals are sent between receivers and satellites and require line of sight for accurate results. Vendors collect location information through hardware installed in a vehicle or data collected through a downloaded smartphone app. GPS pings are frequent and provide a level of detail that allows individual traces to be identified. Currently applications of GPS include network construction, travel time, origin-destination, and traffic counts. All applications have seen significant success except for traffic counts which determined that GPS played an important role in improving traffic count model accuracy. Future work needs to be implemented to derive accurate traffic counts. GPS is an excellent source for obtaining detailed traces that represent true O-D. GPS data is accurate and reliable. Fleet vehicles are often used as a foundation for GPS data sources which can allow differentiation of vehicle modes. Truck or taxi data can be isolated and analyzed separate from the entire population which makes GPS good for targeted studies. GPS often has lower penetration rates and is prone to sample bias because of its strong fleet vehicle representation. The ability to track true O-D presents potential privacy issues because home and work locations can be inferred; all trips are potentially included in the data, but other trip types do not pose as great a privacy threat. Most vendors de-identify data to TAZs to hide an individual's identity. In general, GPS is best suited for vehicles and has not been extensively used for tracking active transportation or transit modes.

CHAPTER 4 - MOBILE DEVICE DATA

4.1 Overview

A large proportion of individuals in modern society carry mobile electronic devices at all times as a matter of habit, utility, and convenience. These devices – principally mobile phones but also smart watches, fitness trackers, music players, and others – connect to data networks in order to provide services to their users. A number of such services require the device to know its location in space: mapping and navigation services, weather updates, and more. Even before the proliferation of these modern services, cellular phones send data to and from towers that are located in space. Over the last ten years, private firms have developed techniques to gather, aggregate, and resell data from cellular towers and location-based services (LBS). MDD is then aggregated, processed and sold to transportation agencies to be used in planning exercises.

In this chapter we review the literature surrounding the use of MDD in transportation planning. The chapter begins with a discussion of the technologies involved in MDD, including cellular triangulation and LBS. Though the technologies underlying the two types of data are different, the resulting data products are used in the same manner by transportation planning agencies. Further, third-party aggregators have largely stopped reselling cellular CDR data as of summer 2020. Nevertheless, the literature using cellular data still contains lessons that apply to LBS data, so we include these studies in the review that makes up the bulk of the chapter.

4.2 Mobile Device Data Underlying Technologies

As discussed above, two separate technologies underly the mobile device data that is resold by third-party data aggregators: cellular call data records that are geospatially located through cellular towers, and location-based services within a mobile device.

4.2.1 Cellular Call Data Records

Cellular networks are composed of cellular towers and base transceiver stations installed and maintained by each telecom company. Cellular phones connect to service networks by sending radio signals to and from cellular towers. By measuring the signal strength of a device among multiple cellular towers or base transceiver stations, a telecom company is able to triangulate a device's location relative to the towers at each point in time. The location is triangulated for outgoing and incoming text messages, calls, and internet connections. Telecom companies also track location information when a handoff between towers occurs. Handoff to an adjacent cellular tower may occur due to increased call volume on the cellular network or movement of a customer to a new physical location. Occasionally a carrier may randomly scan for a device after periods of prolonged inactivity. These random scans, known as periodic location updates, help carriers to reestablish the location of a customer (Miller, 2017).

Ping rates for CDR data are often dependent on the user placing a call or sending or receiving a text message. Cellular companies only record locations and timestamps when a user utilizes the cellular function of the device, changes location between towers, or the company reestablishes a customer's location through periodic location updates. Since data is collected only periodically, in many cases CDR records have large temporal gaps between recorded locations (Iqbal, 2014). The long inactivity of users may have the effect of producing data with missing location histories which could reduce the data's usability.

Similar to the method of trilateration via GPS discussed in Chapter 3, cellular triangulation accuracy is dependent on the number of cellular towers a device is within range of. In general, cellular triangulation has lower spatial resolution because it relies on a cellular phone's relative location to a cellular tower. A study by Tran (2015) determined that without the supplemental aid of GPS satellites, three cellular towers in a rural location can locate an individual within a ¾ square mile region. There are characteristics of cellular service offerings that can affect this spatial precision. In instances where the nearest cellular tower is busy, the next nearest cellular tower is then used to handle the phone's cellular traffic. This improves service for the customers but degrades the precision of the spatial estimates beyond ¾ square mile in a rural area. Although the author did not specify how much improvement was made in urban environments, an improvement in locational accuracy increased with the shortened distance and increase in number of cellular towers.

In the 1995, The Federal Communications Commission started an initiative for improving emergency location services. The initiative was dubbed enhanced 911 (e911) and eventually inspired Congress to institute legislation which required telecom companies to be able to locate an emergency caller to within 160 ft 67% of the time or 490 ft 95% of the time (Ratti, 2005, Spinney 2003). To do this, an improvement in GPS was made which is referred to as assisted GPS (aGPS) which uses cellular towers' continual connection to GPS satellites to expedite locating individuals. This improvement allowed agencies to locate individuals up to 10 ft in perfect conditions. The increased capacity of carriers to track customers provided a rich data set that could be used to track the unique traces of individuals.

The penetration rate of CDR data is often large because it corelates directly with the commercial market share of a given telecom company. Some of the largest companies have a market share that fluctuates around 30% (Statistica, n.d.). According to Pew Research Center (2019), 96% of adults have a personal cellular phone of which 84% are smartphones. Due to the near universal ownership of cellular devices, a CDR sample can often be approximated as the commercial market share of that CDR provider.

Many companies such as AirSage and StreetLight have taken CDR data and then provided it to researchers for various transportation applications (Miller, 2017; Huntsinger, 2017; Monz, 2019). In the United States, significant backlash regarding privacy of customer records encouraged every major telecom company to cut ties with third-party data vendors by summer 2019 (J. Rosenworcel, personal communication, May 15, 2019). Consequently, most data vendors have made a shift towards LBS data collected from mobile apps among other sources. Further research might be conducted on how telecoms are using aGPS data internally in applications with relevance to transportation planning, but public transportation agencies currently do not obtain position data directly from cellular carriers.

4.2.2 Location-Based Services

Many applications on modern "smart" devices offer services based on the user's location, or location-based services (LBS). These LBS include navigation and mapping applications, weather and news updates, location-based advertising, and many others. The application receives location information from the device's operating system. The device might determine its location by a variety of methods – including GPS, cellular tower triangulation, or connecting to a Wi-Fi network with a known location – depending on the device's specific hardware and user configuration. Wi-Fi is a wireless system that allows individuals to connect to the internet. A Wi-Fi network is comprised of various receivers that can broadcast an internet connection to computers, phones, and tablets. Unlike GPS, which requires a direct line of sight with satellites, Wi-Fi technology has the ability to locate a position underground within feet of the individual, provided there is a Wi-Fi router in that space. Wi-Fi technology can be used to locate a device above and below ground, or on specific floors in a multistory building.

When software developers include LBS in their application, they will typically do so using a software development kit (SDK) (Li, 2017). An SDK provides the tools and dependency libraries to engage LBS in an application, meaning that the developers do not have to re-build all the LBS infrastructure for each new application. The analogy of going to the grocery store for vegetables which you did not grow is very similar to the use or licensing of SDKs. App developers are able to select various SDKs as a foundational infrastructure for an app. An important consequence of this, in the context of passive data, is that when an application requests location data from a device, the device location is revealed not only to the servers for that application, but also through the servers that operate the SDK (Morrison, 2020). Thus, popular SDK developers – such as Cuebiq and ironForge – collect LBS data from a large number of LBS applications. It is estimated that at least one-third of mobile apps actively track customer locations (Montjoye, 2013). More recent projections show that certain advertisement companies have their code embedded in as much as 40% of apps in the Google Play Store (Waddell, 2020). It is worth noting, however, that companies which develop some of the most common LBS-using applications – including applications by Apple, Google, Uber, etc. – typically have the resources and motivation to develop their own LBS software and services rather than rely on third-party developers. In doing this, these large companies can develop their own database of device location data, though these companies do not typically sell this data to LBS aggregators.

If a device – and the apps inside it – can obtain location data from Wi-Fi access points, the Wi-Fi access points can also see which phones are nearby. When Wi-Fi in a smart device is activated, the device regularly searches for routers that it can join by sending out a probe request (Freudiger, 2015). These requests might occur up to 55 times per minute and transfer the device MAC address to the router even if the device does not establish a connection to the network (Matte et al., 2017; Li et al., 2017). Large institutions and telecom companies that operate networks of Wi-Fi access points can therefore trace individual devices with some degree of accuracy.

Transportation agencies can purchase location information directly from the SDK developers and Wi-Fi positioning system operators, but more often agencies will buy from third-party aggregators such as StreetLight Data or AirSage. These aggregators clean and process the raw location, time, and device points into forms that transportation agencies can most effectively use. In the remainder of this chapter, we discuss specific applications of these various technologies, though the studies we discuss include a mixture of raw and pre-processed data obtained through third-party aggregators.

4.2.2.1 Social Media Applications

One additional technology bears noting, though it is mostly irrelevant at the writing of this report. Social media applications frequently contain tools whereby users can geolocate their posts, providing a mechanism for a variety of transportation-related studies. Applications where users place themselves at a business or attraction could be used to study trip attraction patterns (Gao 2017; Jin et al. 2014; Yang et al. 2014; Jin et al 2013;); networks where users place photographs might be used to study road conditions or land use (Yan et al. 2019; Sun et al. 2012), and determining home and work location from repeated location observations (Schlieder et al. 2010; Steiger et al. 2016; Huang & Wong 2016; Huang & Wong 2015; Huang 2014;). This kind of data has high and intuitive sources of bias: Users are most likely to post activities that are most interesting to their friends and followers, and the sociodemographic makeup of the most frequent users is unlikely to represent the wider community (Jin et al. 2014; Yang et al. 2015). The ephemeral nature of many social media platforms also calls the sustainability of this data source into question. While potentially useful as a behavior data source in particular research studies, it is unlikely that transportation agencies will be able to use social media platforms in any ongoing projects.

4.3 Transportation Studies Using Mobile Device Data

Mobile device data – including data derived from both cellular CDR and LBS data – has been used extensively for many applications: origin-destination studies, trip attractions, volume estimations, and mode differentiation. Table 3 summarizes several studies presented in this chapter.

Table 3 Mobile Device Data Studies

Author	Year	Success	Notes
Origin Destination			
Caceres et al.	2007	-	Transit surveys
Bernardin et al.	2017	Yes	Long-distance travel demand model
Yin et al.	2017	Yes	Activity-based travel demand model
Huntsinger et al.	2017	Yes	External trip model
Kressner et al.	2017	Yes	Synthetic populations; Travel Demand modeling
Zalewski et al.	2019	Yes	Travel demand modeling / Prioritization of traffic
Trip Attractions			
Friedrich et al.	2010	Yes	Impact analysis (level of service)
McCahill et al.	2017	Yes	Last mile traffic demand
Elkind	2018	-	Impact Analysis (vehicle miles traveled)
Fehr & Peers	2020	-	Impact Analysis (vehicle miles traveled)
Volume Estimations			
Gao et al.	2013	Partial	Volume counts from parallel traces
Turner et al.	2017	No	Volume counts from travel demand models
Codjoe et al.	2018	Partial	Volume counts from travel demand models
StreetLight Data	2019	Partial	Volume counts from travel demand models
Mode Differentiation			
Gao et al.	2013	Partial	Parallel traces & travel times
Bonnetain et al.	2019	Yes	Mapping traces to multi-layered modal maps
Chen et al.	2019	Partial	Above-and-below ground differentiation
StreetLight Data	2019	Yes	Dedicated active transport counters

4.3.1 Origin-Destination Studies

In an origin-destination (O-D) study, the researchers are interested in understanding where trips originate, and where they are going. MDD can be aggregated to generate origin-destination tables at arbitrary geographies or time periods. In these data products, information on the route taken between origin and destination zones is not disclosed. Figure 10 provides a schematic of the data contained in these O-D data sets: The analyst knows the total flows between zones captured by the MDD process, but not any information regarding the path taken. This is for a number of reasons, including privacy protection and also an irregular ping rate relative to GPS data. The remainder of this section will explain how MDD data has been used in various O-D applications.

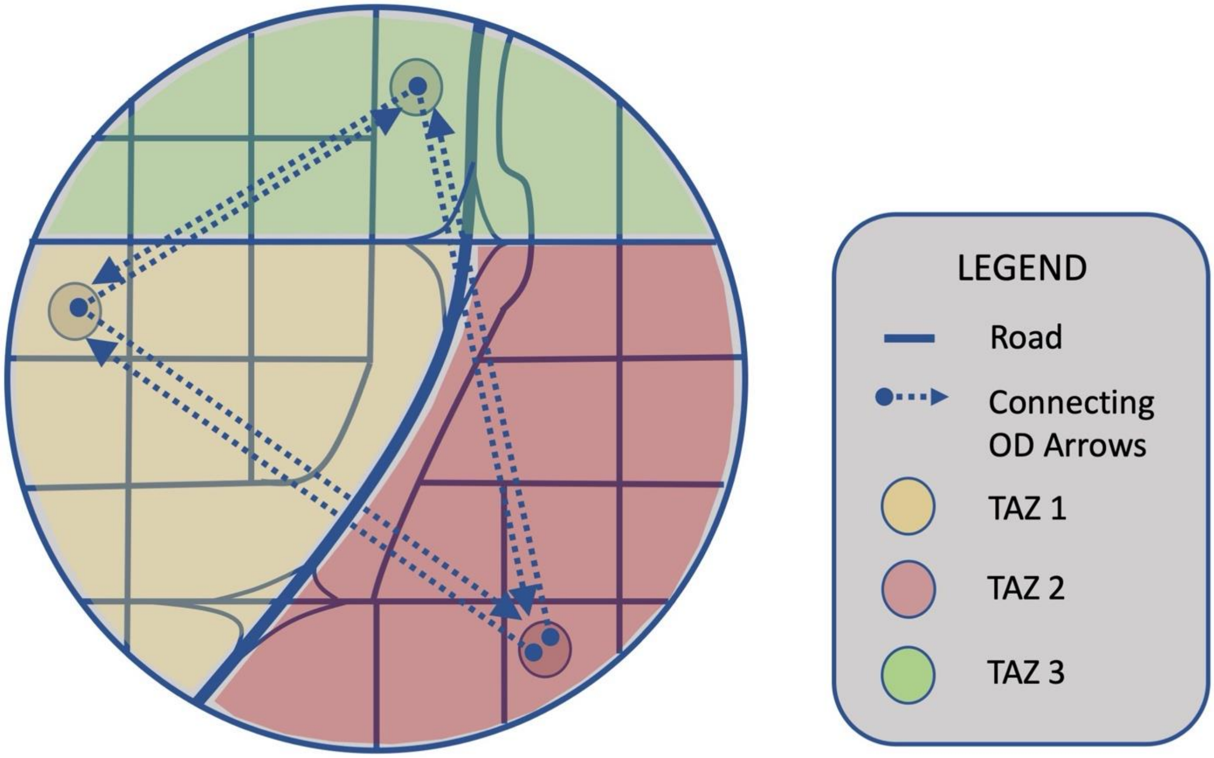

Figure 10. Origin-destination records as represented in mobile device data.

MDD data is frequently used to supplement travel surveys and travel demand models, above and beyond efforts to improve travel surveys by delivering them through location-enabled smartphone applications. Various travel demand modeling tools are created to assist transportation agencies in planning future communities as well as assessing current infrastructure. These models are typically estimated and calibrated from household travel surveys. These travel surveys typically are infrequent, expensive, have small sample sizes, contain inherent bias related to the profiles of survey respondents, and have low response rates overall (Yin, 2017). It is now effectively standard practice to supplement household travel surveys with passive O-D tables derived from MDD when developing regional and statewide travel demand models.

A major limitation in using O-D tables for travel demand analysis is that it is often difficult to know behavioral and sociodemographic details of each trip recorded in the aggregated tables.

One area where this is less of a limitation is in developing external trip models, or submodels that generate and distribute trips that are inbound to the region, outbound from it, or passing through without stopping. In these submodels, the trip purpose and sociodemographic makeup of the trip maker is usually not considered in any meaningful way. The primary consideration is the spatial distribution of the trip origin and destinations, which aggregate MDD can readily provide. As such, a number of studies have shown how O-D tables derived from MDD can be used to generate and inform external trip models. Huntsinger et al. (2017) used CDR data as an input to create external trip models for Asheville, North Carolina, and found that results favored well with comparable travel surveys. In another study on behalf of the Ohio DOT, Miller et al. (2017) compared external trip models developed using multiple data sets, including O-D tables derived from both CDR and LBS. The study emphasized the importance of cleaning and reweighting the O-D tables based on observed highway counts, to correct for potential bias and missing observations in the passive data.

A similar situation to external trips exists with long-distance trips, such as are often found in a statewide travel demand model. Bernardin et al. (2017) used CDR data to update a statewide long-distance travel demand model for Tennessee. They found that long-distance trips were overrepresented relative to short trips in the aggregate data when compared against a household travel survey. Despite this bias, the travel demand model calibrated to the CDR data compared favorably with previous methods and were within 1.5% of observed trips with the exception of one district in their study region. It was concluded that despite inherent bias, CDR data added considerable value toward travel demand models. Other states that are known to have used MDD tables in their statewide modeling efforts include North Carolina, Virginia, and Oregon.

According to personal correspondence with Vince Bernardin, aggregate O-D tables are commonly used to calibrate the trip-distribution model components of a regional travel demand model. It is unlikely to obtain a household survey sample that can generate statistically valid estimates of the flows between zones or districts; the aggregate O-D data derived from MDD can therefore provide a calibration target in addition to other measures. This target is usually made part of a traditional calibration effort, but in a recent application in Charleston, S.C., the O-D table was implemented into the model itself as a “shadow price;” that is, the destination choice model was configured to self-calibrate to the aggregate O-D data.

To date, O-D tables have been largely used as checks or calibration targets in travel demand models. There are ongoing efforts to use passive data in a more robust way in developing the travel models. Yin et al. (2017) developed a generative activity-based travel model by training a series of machine learning algorithms on a large selection of CDR data in the San Francisco Bay region. The resulting model validated well against a more traditional activity-based model developed from a household travel survey. There are also efforts to develop what are called *synthetic* travel diaries. In these efforts, sociodemographic data and MDD O-D tables are joined using various statistical modeling processes. The resulting product contains the O-D patterns observed in the MDD data, but with additional behavioral and sociodemographic data appended to each trip. The data are synthetic, in that they do not actually contain the travel patterns of any real individuals, but Kressner et al. (2017) found that a synthetic model for Atlanta and Seattle compared favorably with existing surveys and other validation tools. For state departments, synthetic modeling has the potential for allowing agencies to obtain data for smaller regions without expensive travel surveys, standardizing reports, and allow for greater targeting of transportation improvements.

Another way in which O-D tables can be used in wider transportation planning efforts is to identify trip distribution patterns that might not be uncovered in a focused travel survey. Caceres et al. (2007) explained that traditional transit travel surveys only targeted existing users of the transit system and failed to address potential riders. In a case study reported in Zalewski et al. (2019), the regional transit agency in Savannah, Georgia purchased aggregate O-D tables for use in their transit service planning. This agency recognized that an on-board rider survey would only reveal trip patterns that were currently possible using their system and would not capture potentially large trip generators that they did not serve well. After using LBS data, the researchers discovered that large numbers of people were frequenting a mall and large health facility that were not connected to existing transit infrastructure. This discovery allowed the agency to reconsider and replan their service offerings.

4.3.2 Trip Attractions

A special case of an O-D study occurs when an analyst or agency wishes to examine trips inbound to a particular zone or facility, rather than O-D patterns in general. Figure 11 illustrates this particular scenario. There may be various reasons to conduct such a study, and there are many examples in the academic and practical literature. For instance, in constructing a new development, you may want to identify a similar facility in a different location and conduct a traffic impact analysis. By identifying where people are coming from to a specific location, you can use those findings as a forecast for the new development.

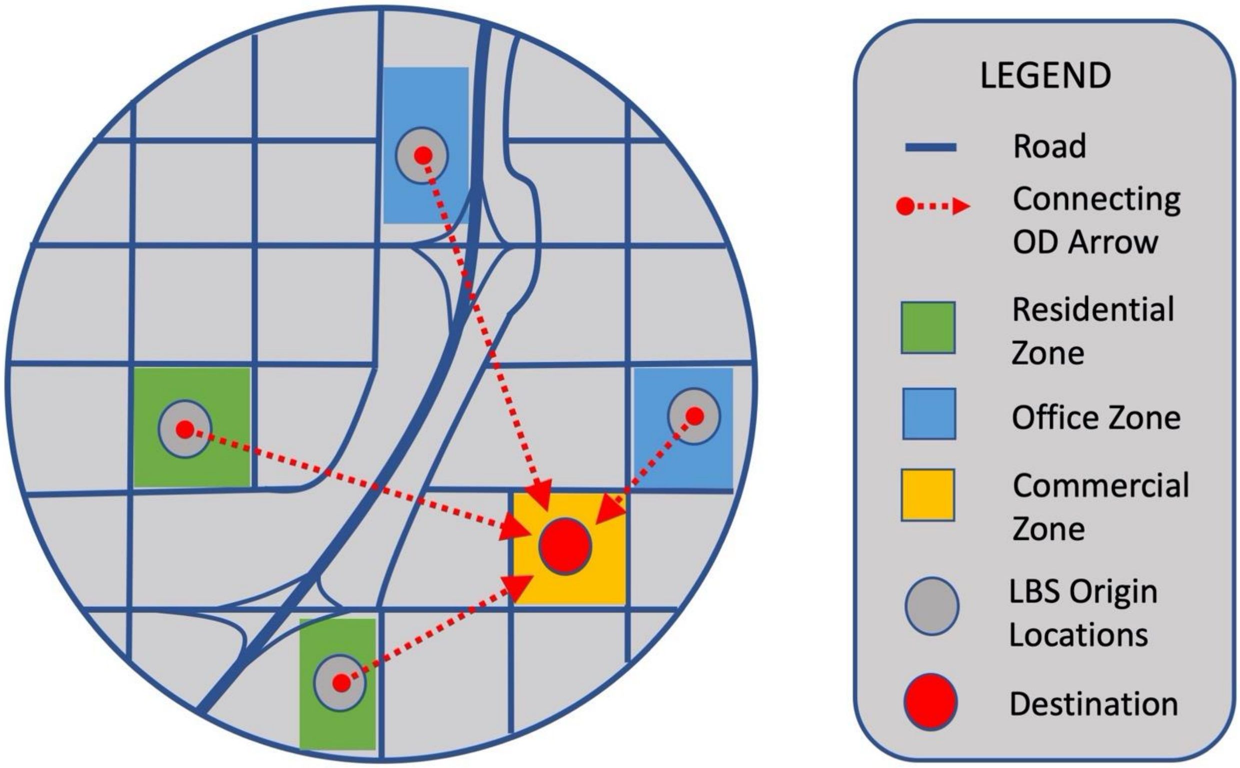

Figure 11. Origin points for trips to a destination facility in MDD.

A major area where trip attraction data is of relevance to transportation planners is in regard to traffic impact studies. A typical traffic impact study will estimate trip attractions to a proposed development using reference materials such as ITE *Trip Generation*, and will then forecast the traffic impacts to local roadways in terms of Level of Service (LOS). In September 2013, the government of California passed Bill-743 (Caltrans, 2020). To adhere to California's climate change reduction goals, the bill replaced LOS with vehicle miles traveled (VMT) as an impact analysis measure. This created two basic problems. The first problem is that *Trip Generation* primarily measures vehicle trips rather than person trips; a VMT-based analysis will show substantively different impacts when there is an appreciable share of non-vehicle trips. The second problem is that the actual origins of the site trips – that is, beyond the specific project area – are not known, thus calculating the average or total trip distance is impossible. Aggregated O-D data can help to resolve both of these problems, as has been shown in a series of recent research and practical efforts (Elkind, 2018; Fehr & Peers, 2020). The availability of these data sets thus enables a more holistic view of traffic impacts than the current LOS-based methodologies.

Beyond traffic impacts, MDD can be used to examine questions of site access for specific land uses. McCahill et al. (2017) used LBS data to identify where last-mile trips near light rail stations were coming from. In this study a major roadway stood between the light rail station and some poorly connected neighborhoods. The region wanted to increase access which would hopefully lead to increased transit ridership. The researchers found that many people were forced to walk through awkward routes to the rail station. The researchers concluded that walking trips were likely to increase if accessibility to the rail stations was better. A post-project analysis also using LBS data could confirm these suspicions. McCahill et al. (2017) admitted that further refinement and validation needed to be performed to solidify the methodology and perform further analysis.

MDD is often sold to firms for help with location analytics: By observing trip origins to existing or competitor facilities, firms can decide where to locate new facilities or expand existing ones. Public planning agencies can follow a similar strategy for identifying locations for parks, libraries and other community resources. Indeed, a number of studies in the park access literature use MDD for exactly this purpose (Monz, et al 2019; Macfarlane et al 2020).

4.3.3 Volume Estimations

Traffic counts are an essential performance measure collected by agencies. These counts are typically collected using temporary pneumatic tubes, radar, software-augmented video systems, or in-pavement loop detectors. In some applications, manual counts are still conducted. Each of these various systems has drawbacks and limitations. Many researchers are beginning to investigate if MDD, particularly LBS, could supplement or replace traditional traffic counting activities. This is a particular concern in rural areas, where traffic infrastructure of all kinds is less available, and where manual counts are marginally more expensive. Figure 12 depicts a simplified scenario where a physical counter detects 30 unique vehicles within a specific time frame. This information is used as a reference for MDD data that is collected at the same time and place as the physical counter data. In this example there is a ratio of 3 MDD traces to 30 actual vehicles (a penetration rate of 10%). In order to use the MDD points, the data must be scaled to represent the entire population. In this case a factor of 10 would be used to boost the sample to represent the entire population. In reality, multiple data sets are used to weight the volume data, including census, demographics, credit spending reports, travel surveys, etc. It is also important to understand that time between MDD locations may be less regular than GPS and so specific traces may not be obtained as easily despite a larger sample of MDD data.

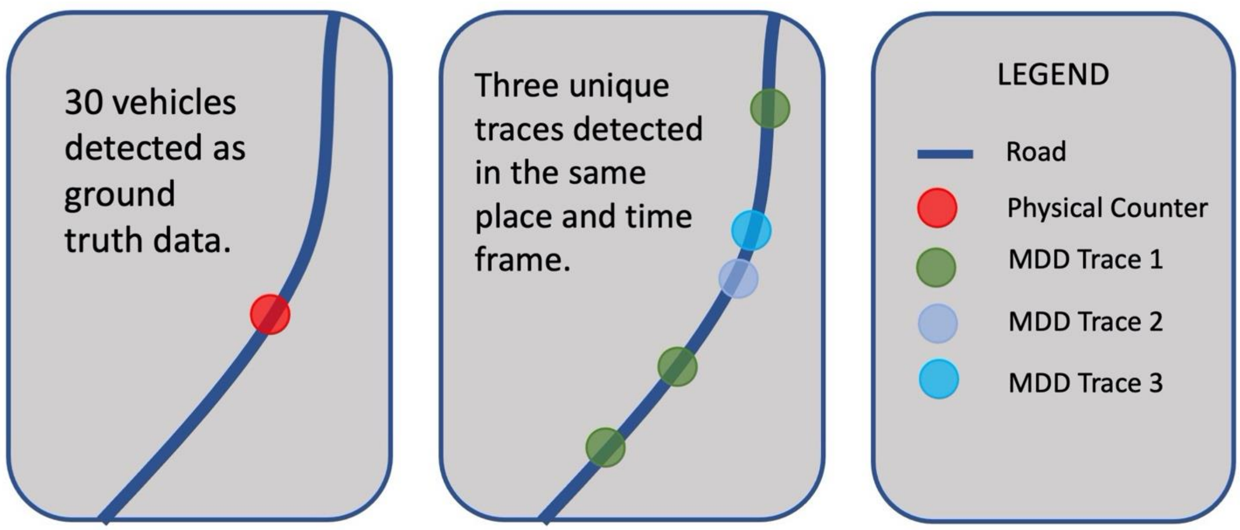

Figure 12. Example of how MDD traffic counts are validated and expanded.

Turner et al. (2017) used LBS and GPS data to create an AADT counter tool derived from a combination of GPS-based apps, other mobile data, and census data. The data was calibrated using permanent and temporary road counters in the state of Minnesota. The research determined that error was substantial – even after removing low-volume roads from the analysis – and refinements to their algorithm would require additional research. The MDD was consistently reporting higher AADT than the permanent counting stations provided by the state. The authors assumed that low sample sizes in low-volume areas contributed significantly to the high errors. At the time CDR was the primary input to the StreetLight Data algorithm and mobile app data was not being used. Their standard errors were approximately 16-40% and needed to be 10% or less. They concluded that tube counts were more accurate than using LBS data.

StreetLight continued to modify their AADT tool introduced in Turner et al. (2017). StreetLight found that the revised tool could find counts in remote areas that traditionally did not have any counts at all. Government organizations without budgets to perform manual counts could use the tool in place of a physical traffic count. The Federal Highway Administration (FHWA) has set a minimum standard from which StreetLight bases its root mean square error (RMSE). StreetLight has achieved an r^2 value of 0.9616 for areas with greater than 5000 AADT. Areas with less than 5000 AADT are not included in this analysis (StreetLight, 2019), and independent verification of these results would be beneficial. As traffic decreases in remote areas the error in their data increases dramatically which was further validated by ODOT (Roll, 2019). They argue that regardless of errors in the data, their methods outperform traditional traffic counts which also have large error due to the short nature of traffic count studies. ODOT also concluded that the use of third-party data would likely result in cost reduction when compared to traditional data collection methods. ODOT confirmed that AADT can be reasonably predicted on roads with greater than 5000 AADT. They also confirmed that data for less traveled roads had high errors and traffic counts were consistently overestimated. ODOT suggested sharing some of their state-collected data to help cater and train the StreetLight party algorithm. They cautioned that only a portion of their data should be used to maintain the ability to conduct independent evaluations in the future.

Rural AADT counts have also been attempted by Citilabs, a software firm. Codjoe et al. (2018) under the Louisiana Department of Transportation used and validated an AADT tool named Streetlytics by Citilabs. This tool works by assigning O-D tables purchased from another third-party vendor to a highway network using a static-user equilibrium-route assignment procedure. Despite the large errors, the researchers determined that federal guidelines were often met while using the Streetlytics software. Similar to StreetLight’s tool, the Citilabs tool successfully provided counts for all roads in Louisiana. They determined that data provided by Streetlytics was comparable to traditional count data which also contains large errors. They determined that the count data was valuable especially for areas where traditional counts may not exist. The agency determined that the data would provide acceptable results while likely reducing the cost of deploying manual counts significantly.

To obtain accurate AADT counts it is important that only one device per vehicle is counted. Some vehicles may have multiple devices which could overpredict AADT counts. It was noted earlier that StreetLight Data seemed to consistently overestimate traffic counts (Roll, 2019). The presence of multiple cell phones could be a cause in these overestimations. Gao et al. (2013), using a simulation, assumed that there is often more than one cellular device in each vehicle. The researchers created an algorithm that clustered cellular traces if they were moving in unison. Using the clustered data and average recorded speeds of vehicles, the researchers were able to differentiate between cars, buses, and trucks. The number of phones in one vehicle was used to identify buses, while average speeds were used to identify cars from trucks. Trucks typically have lower traveling speeds. This method is interesting because it removes potential bias from counting multiple phones in one vehicle. This paper attempts to use mode differentiation to reduce traffic count error. The main issue with the author's methodology was disregarding congestion. The researchers noted that in free-flow traffic, their algorithm was quite accurate but in moments of congestion the tool failed to work because it relied on the relative velocities of traffic. In addition, there was the chance that vehicles traveling parallel down a highway might have the same velocity and consequently be grouped into one vehicle resulting in underestimations. Further work is required to address these issues.

The high usage of free Wi-Fi internet can be used to estimate foot traffic and transit volumes, particularly in institutional settings, though no third-party vendors were identified that sell this data. Similar to Bluetooth, estimated Wi-Fi positions can be recorded as an individual passes a receiver, even if the device does not establish a data connection. The collected MAC addresses and their histories can be aggregated into a series of routes traveled, can identify duration of stay, and can identify the number of people in a controlled setting such as a university or mall (Matte, 2018). In Europe, a university used Wi-Fi to track the movements of students on campus between buildings. They could also determine duration of stay within each building. They found that the technology compared favorably with questionnaire validation surveys (Kalogianni, 2015). In another study, Transport for London launched a study in 2016 that utilized its in-station Wi-Fi network to track O-D information from its underground riders. The study indicated that a third of transit riders were connected to the Wi-Fi network. They also indicated that ridership counts could be estimated to improve transit schedules, mitigate disasters, provide estimated time of arrival to passengers, and increase targeted advertising revenue (Transport for London, 2017).

From current literature, it is apparent that LBS is best used to supplement traditional counts. Traditional counts and LBS counts can work together to give agencies a more detailed image of what is happening on various roadways. Both vendor and traditional counts contain bias and error and combining such data sets limits the impact of erroneous data. (V. Bernardin, personal communication, July 2020)

4.3.4 Mode Differentiation

Traditionally, passive data has been able to identify the traces of motorists with ease but struggled to track the movements of cyclists and pedestrians. The proliferation of cellular phones and MDD presents new opportunities to identify mode-specific detail. Cellular phones have become highly personalized and typically represent a unique individual. Most phones contain dozens of apps that track user location. The large amount of mobile app and CDR data has high probability of representing a diversity of modes. This is a distinguishing feature relative to GPS data, which typically only contains vehicle data, and even then, only a biased sample of mostly fleet vehicles. The challenge, which will be discussed in this section, is how to separate data into distinguishable modes. Figure 13 shows various techniques that have been used to identify specific modes. (a) Shows how travel times between points can help to identify travel time and therefore estimate the mode of unique traces. In this graphic it is assumed that the time intervals are the same for both colored dots. In this case, the green dots represent a slower vehicle because the individual has traveled less distance in the same time as the blue dots. (b) Represents unique traces that are traveling in unison. In some cases, this could identify a multi-passenger vehicle such as a bus. (c) Shows how existing maps may reveal mode. In this image, the blue dots might indicate light rail transit. (d) If underground transit exists, techniques exist to identify subway traffic. (e) Shows how some cities have invested in bicycle counting infrastructure.

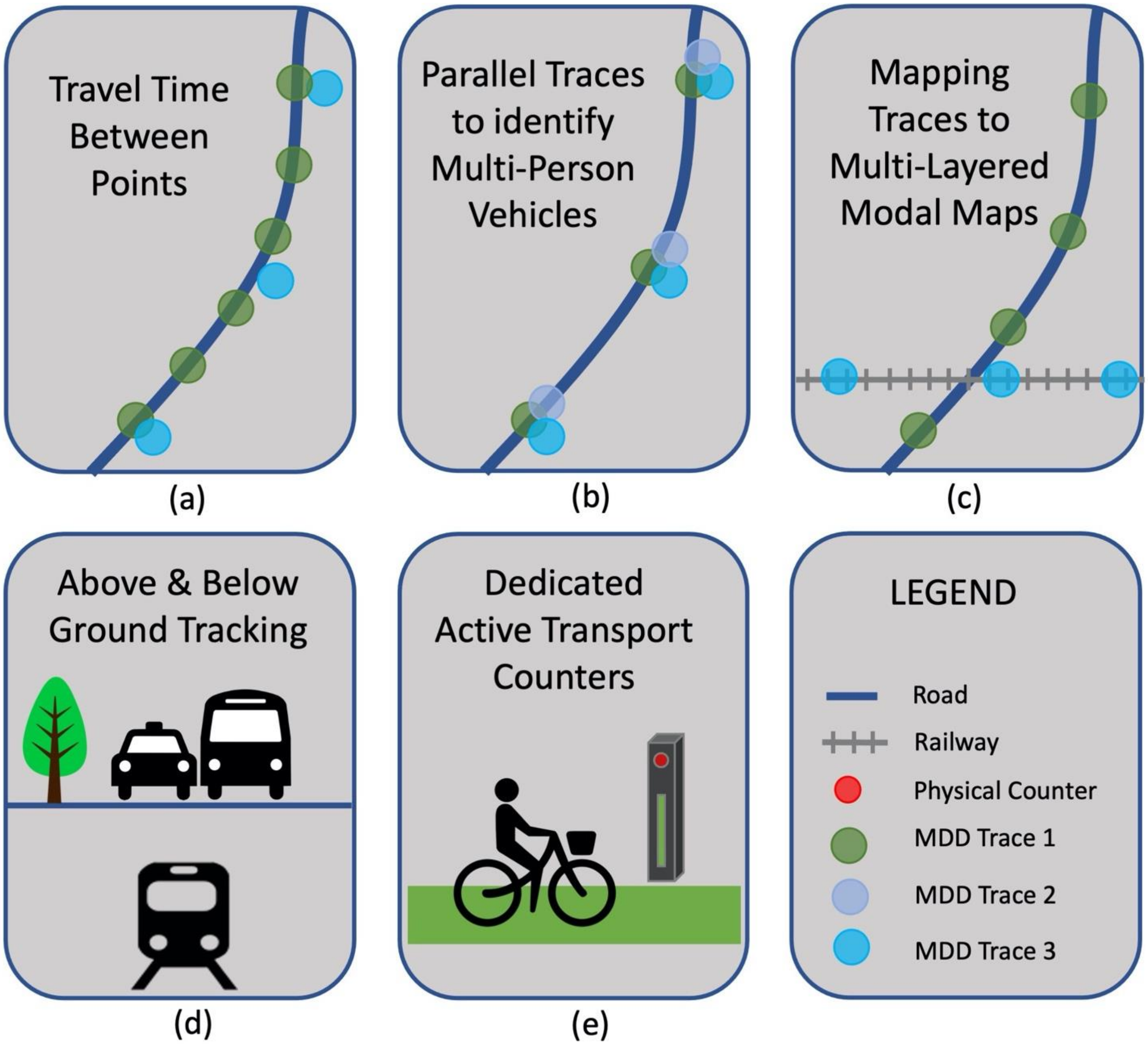

Figure 13. Diagrams showing some of the methods for differentiating mode.

Table 4 Mode Differentiation Techniques

Studies	Methods	Success
Gao et al. (2013)	Parallel traces & travel times	Partial
Bonnetain et al. (2019)	Mapping traces to multi-layered modal maps	Yes
Chen et al. (2019)	Above-and-below ground differentiation	Partial
StreetLight Data (2019)	Dedicated active transport counters	Yes

Multi-passenger vehicles may have parallel traces that could identify mode. Gao et al. (2013) used a clustering approach mixed with travel speeds to differentiate between cars, trucks, and buses. Knowing that buses typically have multiple people, the researcher identified buses by grouping unique cellular traces moving in unison down a highway. The researcher then separated cars and trucks by average speeds assuming that trucks travel at a lesser velocity. The method was successful in free-flow conditions. The method's reliance on travel times caused its failure in congested areas. In addition, there was a chance that vehicles traveling parallel in a platoon-like formation might be falsely grouped as a bus. Further work needed to be done to address these flaws with parallel traces.

Another approach for determining mode differentiation is matching traces to a pre-defined multi-layered modal map like Bonnetain et al. (2019) did in France. The authors made a four-layer map with layers representing surface-level trams, subway, roads, and buses. Data from a major carrier were then map-matched onto the four-layered map. The researchers noted that specific recurring patterns could help to classify each trace. This report offers possible insights into map-matching techniques that could reveal specific modes.

Chen et al. (2019) developed a model that divided modes into above and below ground. This study utilized cellular base transceiver stations which allows a cellular network to connect to users underground. The base transceivers each had a permanent location that was identified as underground and could therefore be separated from above-ground detections. Modes above ground were identified using traditional travel-time data. The researchers were able to identify the underground subway, cars, buses, bicycles, and pedestrians. They found that pedestrians, bicycles, and motorists were overrepresented. Results in this study required further development.

Physical count data for active modes can be mixed with LBS data to create active mode travel demand models. Caltrans has been working with data from the San Francisco area to create active transportation metrics. San Francisco has an extensive bicycle sensor network. Permanent counting towers adjacent to bike paths track the number of cyclists using designated bike lanes in the region. Using the data from these permanent counting stations in San Francisco, Caltrans and StreetLight developed a tool to differentiate modes: walking, cycling, and driving. Findings indicate that location information can identify a rider's location to within 16-65 feet.

They found that penetration rate of cyclists is as low as 1%, as high as 13%, have an average of 4.8% on weekdays, and have a 5.4% average on weekends. Pedestrian penetration rate was as low as 0.5%, as high as 1.5%, averaged 1.31% on weekdays, and averaged 1.02% on weekends. They found that walking presented the largest challenge because of the ambiguity of when a trip began and ended. Findings show that individuals tend to walk around after arriving to a destination such as work. The time it takes for an individual to settle in one location is recognized by the software as part of the overall trip which incorrectly increases the travel time between two specified locations. To expand the data to the entire population, census data was used. The tool has currently been used to identify which projects to prioritize or to identify potential areas for further investigation by crews and other means (StreetLight Data, 2019).

4.4 Analysis of Mobile Device Data as a Data Source

4.4.1 Strengths

MDD data is often attributed with larger sample rates. CDR penetration rate, which represents the approximate market share of a particular carrier, could be around 30% of the total population. As CDR is increasingly being replaced by LBS data, the penetration rates of MDD are less publicized and are often difficult to determine. Replica (sidewalk labs), a company that uses LBS to generate synthetic populations, states that they have a sample of about 5% of the overall population (Bowden, 2018). Another study using GPS provided by StreetLight, indicated that the data sample size represented approximately 23% of the total population (Monz, 2019). From the literature review it appears that sample rates for LBS will vary drastically depending on region and data vendor. Further research should be conducted to determine each vendor’s penetration rate.

It is assumed that MDD data carries less bias because of its universal proliferation. Wireless carriers are chosen arbitrarily, and many apps are nearly universal. This gives MDD a broad and potentially unbiased sample. There is a possibility, however, that data from specific apps could introduce bias in a sample. For example, an app that tracks cyclists may introduce a bias for recreational cyclists because cyclists who commute may not track their activities with an app. Further research should be conducted to determine possible unknown biases in the data set.

Some data vendors update their data set regularly as MDD data becomes available. One such company updates their data every 6 months (StreetLight, 2020). StreetLight pulls data from many sources, many of which are surveys or census data sets that are updated as infrequently as every 10 years. LBS data is being generated continuously and StreetLight uses this data to update their data set every 6 months. Further research should be done to identify if certain data sets which are frequently updated have any added value for traffic engineers.

MDD is collected by several sources which do not require line of sight. Data can be obtained for underground and indoor areas. Using existing infrastructure such as cellular networks and Wi-Fi installations, MDD has a flexibility that other technologies cannot provide. Wi-Fi has the ability to locate an individual indoors and underground at all times when an individual is within range of a Wi-Fi router. In many places Wi-Fi is offered as a free service in exchange for tracking information from that individual. Transport for London used Wi-Fi to identify the O-D of its riders in the London metro system (Transport for London, 2017). In addition, cellular coverage can be extended underground with the use of Base Transceiver Stations making MDD a better location service than GPS in underground scenarios (Chen, 2019)

4.4.2 Weaknesses

MDD data is rarely sold in a raw format. MDD data is generally collected and heavily filtered and processed by proprietary algorithms. The competitive advantage of keeping methodology secret from competing vendors makes it nearly impossible to know how information is collected and processed. The data is owned by the vendor and the ability to access the data is granted to the agency. This gives agencies less control over data but allows agencies to avoid extensive processing and storage of large data sets.

Passive data sets often contain bias. Bernardin et al. (2017) used CDR data and noted that their long-distance travel demand model overrepresented long-distance travelers. One assumption made was that short-distance travelers may be less likely to pull out a cellular device going to places such as a grocery store, while a long-distance traveler is more likely to make a phone call or use cellular service between destinations. Other biases likely exist which have not been explored in detail.

MDD Data tends to work best in larger O-D studies while being less effective in small study regions. For example, if one is searching for O-D between neighboring TAZ's there tends to be a false population generated (Colak 2015). Miller et al. (2017) also indicated that removing some of the TAZ detail may reduce cost as well as increase accuracy in O-D results. Miller et al. (2017) indicates that short trips, which constitute a large portion of daily movements, were missed. They hypothesized that a vehicle's origin and destination may be within the same TAZ which would not be recorded as a trip if models are reducing data accuracy to the nearest TAZ. Further obfuscation of personal information may also lead to blurred location. Friedrich et al. (2010) also indicated that in order to derive useful data, smaller regions required longer aggregation periods. In general, large regions were easier to derive O-D.

CDR has lower spatial resolution as the location is estimated in relation to a fixed tower. Cellular triangulation precision is dependent on the number of cell towers as well as the spacing of those towers in relation to the object. Colak et al. (2015) mentions the "inherent noise contained in the data due to tower-to-tower call balancing." This jump from tower to tower may signal a trip generation when in reality the individual remained stationary. A cellular tower can typically locate individuals to within 330 feet of the tower (Miller, 2017), but locations may be off as much as 980 feet. Tran (2015) noted that without the aid of GPS satellites, three cellular towers can locate an individual within a ¾ square mile. In rural areas where cellular towers are more sparsely located accuracy can be diminished. In instances where the nearest cellular tower is busy, the next nearest cellular tower is then used to locate a cellular phone which can extend the accuracy beyond ¾ square mile. For MDD sources the spatial resolution may also be difficult to isolate. The locational precision of an individual app user may be affected by how the device is connected. If the device is connected to a Wi-Fi network it will log a location based on where the router is. If an app user is driving the location will be recorded using the GPS functions of the phone. These differences in how a location is recorded for the same user may result in truncated precision at different levels. Further investigation should be conducted to see if location precision is truncated differently between apps. The aggregation of data from many individual apps may provide varied levels of precision that may lower the spatial resolution of the collective data set (V. Bernardin, personal communication, July 2020).

MDD often has large temporal gaps. CDR records in particular can have large temporal gaps (Iqbal, 2014). To fill in the temporal gaps found in MDD, vendors often aggregate data from multiple sources over longer time periods. Unlike GPS which often has a set frequency of pings or locations recorded, MDD does not. MDD may contain sporadic spatial gaps with varied times between pings. In many cases the purchased MDD data will only contain origin and destination points and will not include the exact trace of individuals. The lack of information between O-D and the low ping rates often disqualifies its use for real-time applications. Low ping rates can also affect the usefulness of data when unique identifiers are frequently reset. Freidrich et al. (2010) used CDR, but found that their country regulations required unique identifiers on devices to be reset after 24 hours which was not long enough to derive all the information they needed. In the United States, unique identifiers are generally not required to be reset but temporal gaps are very common.

Penetration rates are generally high, but the actual rate is often unspecified. As MDD has evolved more towards information collected from smartphone apps, it has become less obvious where data is coming from. Vendors may collect data from cellular apps, GPS, CDR, census data, credit history reports, etc. It is assumed that vendors with aggregated sources have maintained higher penetration rates than Bluetooth or GPS but further investigation should be conducted to confirm the assumption in this paper. In general, sample rates should be investigated on a vendor-to-vendor basis.

4.5 Summary and Recommendations

Mobile device data, which comes in multiple forms such as CDR or LBS, is collected from cellular phones. CDR are records for billing purposes. Data is tracked by the telecom company when an individual sends or receives texts, calls, or accesses the internet. LBS is collected from various apps. Many of the apps downloaded on smartphones contain embedded code which enables application vendors and location service providers to identify locations and associated times. The location information is aggregated by data vendors who process and sell data to transportation agencies.

LBS has many strengths and weaknesses. The wide proliferation of smartphones makes LBS a rich data set that is suspected to have less bias. Sample rates can be as high as a telecom's commercial market share. In cases where LBS data is sourced from mobile apps, penetration rates have been lower but can still be comparable if not higher than other passive data sources. It is important to understand that LBS ping rates are much more variable than GPS ping rates which typically occur at consistent intervals. LBS has an advantage in underground and indoor settings because Wi-Fi does not require line of sight like GPS technology does. LBS technology has the ability to be as accurate as GPS sources but can have a distribution of accuracies depending on which location technology is being sourced. LBS data may record locations relative to cellular towers, Wi-Fi routers, or the GPS location of the user. Spatial accuracy can be up to 980 feet off of the true location in extreme cases. LBS tends to work best in larger study regions because of the lack of spatial precision and the lack of regular pings. Origin and destination points are often all that are provided because of a lack of specific trace information.

LBS data has been used extensively in traffic demand models with varying success. It is recommended that LBS data be mixed with historical travel survey information for a more complete picture. Much of the research on the use of LBS data is ongoing. Several researchers have failed to use LBS for AADT counts. LBS seems promising for use in mode differentiation. Recently, California has switched from using LOS as an environmental performance measure to a VMT model. LBS seems like a promising technology that could aid agencies in VMT estimations. Further research is needed to determine LBS data's effectiveness in these mentioned applications. In general, it is evident that LBS, if not a standalone data set, is highly valuable as a supplemental data source. The mix of LBS data with other sources provides a more complete picture of the ground truth.

CHAPTER 5 - CONCLUSIONS AND RECOMMENDATIONS

5.1 Overview

In this report, the utility to transportation departments of three passive data sources were explored: Bluetooth, GPS, and mobile device data (MDD) which includes both LBS and CDR as underlying technologies. Passive data exists as a result of other processes which generate data as a byproduct which can be used for various transportation projects. Passive data may be created as a result of individuals connecting to a cellular network, vehicles navigating by way of GPS, the use of cellular phone applications, or financial firms monitoring the credit of potential borrowers. As mentioned in the introduction of this report, a number of firms have developed business models that center around the aggregation and reselling of passively collected data. It is essential for UDOT and other transportation agencies to understand current capabilities of passive data sources and the applications of that data. This paper investigated how peer agencies and academics were using passive data so that UDOT can make informed decisions on how to leverage the data for Department purposes.

Bluetooth is a short-range communications protocol that allows authenticated devices to send limited information between each other. Bluetooth devices are constantly broadcasting a MAC address in attempting to connect to other Bluetooth-enabled devices. Transportation analysts are able to intercept these broadcast MAC addresses as individuals pass by a Bluetooth receiver setup. Bluetooth data is able to obtain travel times and speeds when a unique device is detected at two or more Bluetooth receiver stations. It is important to remember that Bluetooth in unable to determine true origin and destination locations. Transportation analysts are able to set up a cordon, a temporary perimeter around a study area, and obtain timestamps of when an individual enters and exits the cordon perimeter. Bluetooth technology seems well fitted for single-day event mitigation and planning (Rescot, 2011), temporary events such as construction projects (Haseman, 2010), and before-and-after traffic comparisons (Kim et al. 2014). Bluetooth allows agencies to target specific areas and obtain customized data collection because of its easy customizable set up. This makes Bluetooth well suited for obtaining real-time estimates of travel time. Path analysis is possible but difficult without a permanent and extensive setup of Bluetooth receiver hardware (Jackson and Dichev, 2013).

GPS is a well-established technology that uses trilateration of satellites orbiting earth with GPS devices on the surface of the earth. GPS is highly accurate but individual cellphones may introduce errors of up to 16 feet in open sky conditions (Diggelen & Enge, 2015; Tomastik & Mokros, 2017). GPS requires a line of sight which makes it less accurate near tall buildings or natural features. There is a large concentration of GPS data that comes from fleet vehicles which makes it an attractive data source for those trying to isolate vehicle modes such as commercial trucking or fleet taxis. GPS data is sold in varied formats depending on the data vendor. Obtaining raw data gives agencies additional control but may overwhelm agencies because of the amount of processing that is required to remove extraneous information. GPS is well suited for O-D application because data contains true start and end points from which home and work locations can be inferred. This allows data to be used in travel demand models. GPS data also contain regular ping rates or intervals where a timestamp is created which allows transportation analysts to identify exact traces through a transportation network. Average travel speeds between pings are obtainable and sometimes instantaneous speeds are available for each ping. Other successful applications have included road network construction (Biagiani and Eriksson, 2012), travel-time reliability indexing (Stimpancic, 2016; Pinjari, 2014), bottleneck identification (McCormack, 2011; Liao, 2014), duration of stay (Golias, 2012), commodity flow (Zhao, 2020), and truck parking demand (Diaz-Corro, 2019).

MDD data comes from two main sources: CDR and LBS. LBS data can be collected through smartphone applications and connections to Wi-Fi networks. Most commonly, LBS data is collected through smartphone applications which use SDK packages provided to app developers as building blocks for their apps. Third-party companies allow app developers to use SDKs which can send user data back to the creator of the SDK. The companies that create and own SDKs can then use the data for a variety of purposes including selling aggregated data to transportation analysis companies. CDR data is generated when individuals call, text, or connect to the internet through their cellular network. CDR data can also be collected when a signal jumps to a different cell tower or the company randomly searches for the device. In general, MDD has well defined start and end points but often lacks intermediate trip data. For this reason, LBS data – as usually provided by third-party vendors to transportation agencies – often only contains the start and end points of a trip. MDD has lower spatial accuracy than that of Bluetooth or GPS, but the sample size tends to be higher because of its multiple collection sources. MDD data is best used for input in travel demand models and can identify trip attractions. MDD has been used with mixed results to determine traffic counts and mode differentiation.

5.2 Recommendations and Catalog

This report contains a number of specific recommendations for UDOT in strategizing, obtaining, and using third-party data sets for transportation planning and analysis. These recommendations are summarized here as follows:

1. UDOT staff need to be aware of the source technologies of third-party data sets, along with their strengths and weaknesses.
2. UDOT should develop consistent data validation routines to identify the inherent accuracy of purchased data products, and regularly evaluate new data purchases.
3. UDOT should investigate using permanent Bluetooth receivers to measure travel times between points on key corridors. This may be particularly useful in places where GPS and cellular reception might be unreliable, such as in canyons.
4. UDOT should not rely on permanent or temporary Bluetooth receivers to analyze trip origins, destinations, routes, or volumes outside of small, well-defined cordon studies or institutional settings.
5. UDOT should use GPS data to develop a more complete picture of freight movements in and through the state.
6. UDOT should continue to investigate the use and applications of aggregate GPS data in determining speeds on roadways.
7. UDOT should work to develop an integrated transportation planning approach that relies on the relative strengths of household travel surveys as well as third-party origin-destination data derived from LBS.
8. UDOT should avoid relying on a single vendor or data source for any of its analyses, and instead develop processes that make use of multiple data inputs.

Table 5 lists the major strengths and weaknesses of the three passive data sources synthesized in this report. For more detail on the strength and weaknesses of each individual technology reference the chapter for each technology. The last column shows a few of the major companies that sell passive data for transportation purposes.

Table 5 Major Strengths and Weaknesses of Passive Data and Major Data Companies

Data Source	Strengths	Weaknesses	Major Companies
Bluetooth	- Temporary event data collection - Real-time estimates of travel time - Customized cordon studies - Accurate relative to receivers	- Obtaining traces through a cordon is difficult - Hardware setup required - Minor timestamping issues	- Blyncsy - TrafficCast (Bluetoad)
GPS	- High accuracy - True O-D data - Regular ping rates and intermediate trace data - Travel times from average speeds as well as possible instantaneous speeds available - Fleet vehicle origins allow for targeted mode data sets such as taxis or commercial trucking	- Low penetration rates - Cell phones may introduce up to 16 feet error in clear conditions - Strong bias from fleet vehicle origins - Technology requires direct line of sight	- ATRI - HERE Technologies - INRIX - TOM TOM
MDD	- True O-D data - Trip attractions can be identified and used in travel demand models - Large sample sizes - CDR has low bias, it is unknown if mobile application data has significant bias - More opportunity for tracking active modes and mode differentiation	- Lack of intermediate trace data. Data often only contains start and end points - Lower spatial accuracy - Data works best in large O-D areas.	- AirSage - Citilabs - Replica – Sidewalk Labs - StreetLight Data

5.3 Summary

Passive data is a growing market that has many transportation-related applications. This synthesis report has confirmed that passive data can be a valuable asset to transportation agencies. Passive data represents an alternative data source and perspective in comparison to traditional methodologies such as manual counts and travel surveys. After reviewing current academic literature and interviewing industry professionals who have firsthand experience with the passive data sources in this report, it is clear that there is no singular data source that can replace traditional manual data collection methods. Each passive data source has strengths and weaknesses that make it valuable for certain situations. It is evident that mobile phone use is continually increasing with increases in mobile application production and use. Bluetooth devices are increasingly common in motor vehicles, headsets, cellular phones, and other consumer electronics. It is unclear if GPS data from physical devices is growing or if data is shifting towards navigation application on mobile phones. Regardless, GPS equipment is common in fleet vehicles and provides DOTs with important data sets for specific modes like taxis and commercial trucks. Literature indicates that these technologies are increasingly common and often successful in various transportation studies.

Passive data works hand in hand with manually collected data to expose the ground truth. Passive data sets – like all sampled data – represent a portion of the entire population and therefore contain inherent bias. It is nearly impossible for a data set to perfectly represent actual conditions. Passive data sets contain bias such as overrepresenting long-haul trucking (Bernardin, 2014) that when scaled to represent the total population may further amplify the bias. Although traditional manual data collection methods can be entirely replaced with passive data, the combination of passive and manual data sources may reduce bias and error. Several researchers in literature have noted how RMSE goes down when passive data is input into various travel models. One interviewed researcher indicated that it is unlikely that passive data will replace traffic counts completely, but that passive data has the potential of significantly reducing the regularity or scale of traditional counts. The researcher went on to state that passive data sources combined with traditional counting methods provide DOTs with a closer representation of the truth (V. Bernardin, personal communications, July 2020). This trend is likely to occur across many transportation applications that use manual data collection. Passive data could be used as a preliminary analysis tool to identify areas that need further attention. Using passive data to identify areas of interest will help agencies better allocate human resources. In conclusion, passive data is a quick and valuable tool that can help agencies obtain results closer to the ground truth.

REFERENCES

Altintasi, O., Tuydes-Yaman, H., & Tuncay, K. (2016). Detection of urban traffic patterns from Floating Car Data (FCD). *19th EURO Working Group on Transportation Meeting, EWGT2016*, 382–391. https://doi.org/10.1016/j.trpro.2017.03.057

Asakura, Y., Kusakabe, T., Long, N. X., & Ushiki, T. (2015). Incident Detection Methods Using Probe Vehicles with On-board GPS Equipment. *Transportation Research Procedia*, *6*(June 2014), 17–27. https://doi.org/10.1016/j.trpro.2015.03.003

Ban, X. J., Hao, P., & Sun, Z. (2011). Real time queue length estimation for signalized intersections using travel times from mobile sensors. *Transportation Research Part C: Emerging Technologies*, *19*(6), 1133–1156. https://doi.org/10.1016/j.trc.2011.01.002

Barceló, J., Montero, L., Marqués, L., & Carmona, C. (2010). Travel Time Forecasting and Dynamic OD Estimation in Freeways Based on Bluetooth Traffic Monitoring. *Transportation Research*, (November 2009), 1–20. https://doi.org/10.3141/2175-03

Bastani, F., He, S., Jagwani, S., Park, E., Abbar, S., Alizadeh, M., … Sadeghi, M. A. (2019). *Inferring and Improving Street Maps with Data-Driven Automation*. Retrieved from http://arxiv.org/abs/1910.04869

Bathaee, N., Mohseni, A., Park, S. J., Porter, J. D., & Kim, D. S. (2018). A cluster analysis approach for differentiating transportation modes using Bluetooth sensor data. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *22*(4), 353–364. https://doi.org/10.1080/15472450.2018.1457444

Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., & Giovannini, L. (2010). Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *Journal of Statistical Mechanics: Theory and Experiment*, *2010*(5), 1–17. https://doi.org/10.1088/1742-5468/2010/05/P05001

Bernardin, V. L., Ferdous, N., Sadrsadat, H., Trevino, S., & Chen, C. C. (2017). Integration of national long-distance passenger travel demand model with tennessee statewide model and calibration to big data. *Transportation Research Record*, *2653*(1), 75–81. https://doi.org/10.3141/2653-09

Bernardin, V. L., Trevino, S., & Short, J. B. (2014). *Expanding Truck GPS-based Passive Origin-destination Data In Iowa And Tennessee*. Retrieved from http://www.ghbook.ir/index.php?name=فرهنگ و رسانه های نوین&option=com_dbook&task=readonline&book_id=13650&page=73&chkhashk=ED9C9491B4&Itemid=218&lang=fa&tmpl=component

Biagioni, J., & Eriksson, J. (2012). Inferring road maps from global positioning system traces. *Transportation Research Record*, (2291), 61–71. https://doi.org/10.3141/2291-08

Blogg, M., Semler, C., Hingorani, M., & Troutbeck, R. (2010). Travel time and origin-destination data collection using Bluetooth MAC address readers. *ATRF 2010: 33rd Australasian Transport Research Forum*, (October), 1–15.

Bluetooth SIG. (n.d.) *The story behind how Bluetooth got its name*. Bluetooth. Retrieved May 26, 2020. from https://www.bluetooth.com/about-us/bluetooth-origin/

Bluetooth. In *Wikipedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=Bluetooth&oldid=958983432

Bonnetain, L., Furno, A., Krug, J., & Faouzi, N. E. El. (2019). Can We Map-Match Individual Cellular Network Signaling Trajectories in Urban Environments? Data-Driven Study. *Transportation Research Record*, *2673*(7), 74–88. https://doi.org/10.1177/0361198119847472

Bowden, N. (2018, April 6). *Introducing Replica, a next-generation urban planning tool* Medium. https://medium.com/sidewalk-talk/introducing-replica-a-next-generation-urban-planning-tool-1b7425222e9e

Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Deriving origin - destination data from a mobile phone network. *IET Intelligent Transport Systems*, *2*(2), 132. https://doi.org/10.1049/iet-its

Caltrans. n.d. *Senate Bill (SB) 743 Implementation* https://dot.ca.gov/programs/transportation-planning/office-of-smart-mobility-climate-change/sb-743

Carpenter, C., Fowler, M., & Adler, T. (2012). Generating route-specific origin-destination tables using bluetooth technology. *Transportation Research Record*, (2308), 96–102. https://doi.org/10.3141/2308-10

Chen, C., Lu, C., Huang, Q., Yang, Q., Gunopulos, D., & Guibas, L. (2016). City-scale map creation and updating using GPS collections. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 1465–1474. https://doi.org/10.1145/2939672.2939833

Chen, X., Wan, X., Li, Q., Ding, F., McCarthy, C., Cheng, Y., & Ran, B. (2019). Trip-Chain-Based Travel-Mode-Shares-Driven Framework using Cellular Signaling Data and Web-Based Mapping Service Data. *Transportation Research Record*, *2673*(3), 51–64. https://doi.org/10.1177/0361198119834006

Chitturi, M. V., Shaw, J. W., Campbell, J. R., & Noyce, D. A. (2014). Validation of Origin–Destination Data from Bluetooth Reidentification and Aerial Observation. *Transportation Research Record: Journal of the Transportation Research Board*, *2430*(1), 116–123. https://doi.org/10.3141/2430-12

Codjoe, J., Ashley, G., & Saunders, W. (2018). *Evaluating Cell Phone Data for AADT Estimation* (Vol. 70808). Retrieved from http://www.ltrc.lsu.edu/downloads.html

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., & González, M. C. (2015). Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities. *Transportation Research Record*, *2526*, 126–135. https://doi.org/10.3141/2526-14

Cross, J. (2018, March 14). *Bluetooth 5 FAQ: Everything you need to know*. Macworld. https://www.macworld.com/article/3262664/bluetooth-5-faq-everything-you-need-to-know.html

Davies, J., Beresford, A., & Hopper, A. (2006). Scalable, Distributed, Real-Time Map Generation. *IEEE Pervasive Computing*, *5*, 47–54.

Day, C. M., Haseman, R., Premachandra, H., Brennan, T. M., Wasson, J. S., Sturdevant, J. R., & Bullock, D. M. (2010). Evaluation of arterial signal coordination: Methodologies for visualizing high-resolution event data and measuring travel time. *Transportation Research Record*, (2192), 37–49. https://doi.org/10.3141/2192-04

Diaz-Corro, K., Akter, T., & Hernandez, S. (2019). Comparison of Overnight Truck Parking Counts with GPS-Derived Counts for Truck Parking Facility Utilization Analysis. *Transportation Research Record*, *2673*(8), 377–387. https://doi.org/10.1177/0361198119843851

Elkind, E. N., Lamm, T., & Prather, E. (2018). *Implementing SB 743 An Analysis of Vehicle Miles Traveled Banking and Exchange Frameworks October 2018*. (October). https://doi.org/10.7922/G2MK6B22

Fehr & Peers. (2020). *VMT Mitigation Through Fees, Banks, & Exchanges: Understanding New Mitigation Approaches*.

Findings, E. (2017). *GPS (GNSS) accuracy of Lenovo Phab 2 Pro under conditions of windthrown area*. (June). https://doi.org/10.13140/RG.2.2.31978.93123

Freudiger, J. (2015). How Talkative is Your Mobile Device? An Experimental Study of Wi-Fi Probe Requests. *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. https://doi.org/10.1145/2766498.2766517

Friedrich, M., Immisch, K., Jehlicka, P., Otterstätter, T., & Schlaich, J. (2010). Generating origin-destination matrices from mobile phone trajectories. *Transportation Research Record*, (2196), 93–101. https://doi.org/10.3141/2196-10

Gao, H., & Liu, F. (2013). Estimating freeway traffic measures from mobile phone location data. *European Journal of Operational Research*, *229*(1), 252–260. https://doi.org/10.1016/j.ejor.2013.02.044

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, *21*(3), 446–467. https://doi.org/10.1111/tgis.12289

Gately, C. K., Hutyra, L. R., Peterson, S., & Sue Wing, I. (2017). Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone GPS data. *Environmental Pollution*, *229*, 496–504. https://doi.org/10.1016/j.envpol.2017.05.091

General Services Administration. Federal Agency Telematics Experiences August 2014. GSA.gov. Retrieved May 26, 2020. From https://www.gsa.gov/cdnstatic/Agency_Telematics_2014.pdf

GIS Geography. (2020, February 24). *Selective Availability in Global Positioning System (GPS)*. GIS Geography. https://gisgeography.com/selective-availability-gps/

Golias, M., Dobbins, J., Short, J., & Johnson, Z. (2012). *GPS / GIS Analysis of Tennessee Truck Trips*.

Goodall, N. (2012). Microscopic Estimation of Freeway Vehicle Positions Using Mobile Sensors. *Transportation Research Board, 91st Annual Meeting*, (January), 1–14.

Haque, K., Mishra, S., Paleti, R., Golias, M. M., Sarker, A. A., & Pujats, K. (2017). Truck Parking utilization analysis using GPS data. *Journal of Transportation Engineering Part A: Systems*, *143*(9). https://doi.org/10.1061/JTEPBS.0000073

Haseman, R. J., Wasson, J. S., & Bullock, D. M. (2010). Real-time measurement of travel time delay in work zones and evaluation metrics using bluetooth probe tracking. *Transportation Research Record*, (2169), 40–53. https://doi.org/10.3141/2169-05

Here Technologies. (n.d.) *Here Maps*. Retrieved from https://www.here.com/platform/mapping/map-data

Hollander, D. (2019, March 27). *How AoA & AoD Changed the Direction of Bluetooth Location Services.* Bluetooth. https://www.bluetooth.com/blog/new-aoa-aod-bluetooth-capabilities/

Huang, H., Gartner, G., Krisp, J. M., Raubal, M., & Van de Weghe, N. (2018). Location based services: ongoing evolution and research agenda. *Journal of Location Based Services*, *12*(2), 63–93. https://doi.org/10.1080/17489725.2018.1508763

Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate? - A GIS approach for user location inference. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2014 - Held in Conjunction with the 22nd ACM SIGSPATIAL GIS 2014*, 1–8. https://doi.org/10.1145/2755492.2755494

Huang, Q., & Wong, D. W. S. (2015). Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. *Annals of the Association of*

American Geographers, *105*(6), 1179–1197. https://doi.org/10.1080/00045608.2015.1081120

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*(9), 1873–1898. https://doi.org/10.1080/13658816.2016.1145225

Huntsinger, L. F. (2017). The Lure of Big Data: Evaluating the Efficacy of Mobile Phone Data for Travel Model Validation. *Compendium of Transportation Research Board 2017 Annual Meeting*, *9*(January), 1–16.

ITE, Trip Generation Manual, 10th Edition

Ikeda, Y., & Inoue, M. (2016). An Evacuation Route Planning for Safety Route Guidance System after Natural Disaster Using Multi-Objective Genetic Algorithm. *Procedia Computer Science*, *96*, 1323–1331. https://doi.org/10.1016/j.procs.2016.08.177

Iqbal, S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Jackson, P., & Tony, D. (2014). USE OF BLUETOOTH TRACKING TECHNOLOGY IN TRANSPORT. *European Transport Conference 2014*, 1–11.

Jin, Jing; Yang, Fan; Cebelak, Meredith; Ran, Bin; Walton, C. M., Jin, P. J., Yang, F., Cebelak, M., Ran, B., & Walton, C. M. (2013). Urban Travel Demand Analysis for Austin, Texas, Using Location-Based Social Networking Data. *Transportation Research Board 92nd Annual Meeting*, 1–17. Retrieved from http://trid.trb.org/view.aspx?id=1241465

Jin, P. J., Cebelak, M., Yang, F., Zhang, J., Walton, C. M., & Ran, B. (2014). Location-Based Social Networking Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2430*(1), 72–82. https://doi.org/10.3141/2430-08

Kalogianni, E., Sileryte, R., Lam, M., Zhou, K., Van der Ham, M., Van der Spek, S. C., & Verbree, E. (2015). Passive WiFi monitoring of the rhythm of the campus. *Proceedings of The 18th AGILE International Conference on Geographic Information Science; Geographics Information Science as an Enabler of Smarter Cities and Communities, Lisboa (Portugal), June 9-14, 2015; Authors Version*, (June). Retrieved from http://resolver.tudelft.nl/uuid:35734ce7-a9b4-459e-baf2-1f9f0a2391b0#.Vr3_zBgdIYc.mendeley

Kamali, M. (2015). *Development of Truck Route Choice Data Using Truck GPS* (University of South Florida). Retrieved from http://scholarcommons.usf.edu/etd/5968/

Kamali, M., Ermagun, A., Viswanathan, K., & Pinjari, A. R. (2016). Deriving truck route choice from large GPS data streams. *Transportation Research Record*, *2563*(813), 62–70. https://doi.org/10.3141/2563-10

Kan, Z., Tang, L., Kwan, M. P., & Zhang, X. (2018). Estimating vehicle fuel consumption and emissions using GPS big data. *International Journal of Environmental Research and Public Health*, *15*(4), 1–23. https://doi.org/10.3390/ijerph15040566

Karr, S. (2016, April 27). *Consumer Survey Highlights Necessity of Bluetooth Technology*. Bluetooth. https://www.bluetooth.com/blog/consumer-survey-highlights-necessity-of-bluetooth-technology/.

Kattan, L., & Saidi, S. (2012). Journal of Advanced Transportation. *Journal of Advanced Transportation*, *47*(1), 25–26. https://doi.org/10.1002/atr.1187

Kim, D. S., Porter, J. D., Park, S., Saeedi, A., Mohseni, A., Bathaee, N., & Nelson, M. (2014). *Bluetooth Data Collection System for Planning and Arterial Management*. Retrieved from http://www.oregon.gov/ODOT/TD/TP_RES/docs/Reports/2014/SPR_757_Final_Bluetooth.pdf

Kressner, J. D. (2017). Synthetic Household Travel Data Using Consumer and Mobile Phone Data IDEA Program Final Report The National Academies. In *Transportation Research Board*. Retrieved from https://pdfs.semanticscholar.org/0925/61a50d8cf25798cd5e10ed9906f63bfc7780.pdf

Kurkcu, A., & Ozbay, K. (2017). Estimating pedestrian densities, wait times, and flows with wi-fi and bluetooth sensors. *Transportation Research Record*, *2644*(1), 72–82. https://doi.org/10.3141/2644-09

Lehmann, A., & Gross, A. (2017). Towards vehicle emission estimation from smartphone sensors. *Proceedings - 18th IEEE International Conference on Mobile Data Management, MDM 2017*, 154–163. https://doi.org/10.1109/MDM.2017.29

Li, F., Wang, X., Niu, B., Li, H., Li, C., & Chen, L. (2017). TrackU: Exploiting User's Mobility Behavior via WiFi List. *2017 IEEE Global Communications Conference, GLOBECOM 2017 - Proceedings*, *2018-January*, 1–6. https://doi.org/10.1109/GLOCOM.2017.8254030

Liao, C.-F. (2014). *Using Truck GPS Data for Freight Performance Analysis in the Twin Citi es Metro Area*. Retrieved from http://www.dot.state.mn.us/research/TS/2014/201414.pdf

Liu, H., Chen, X., Wang, Y., & Han, S. (2013). Vehicle emission and near-road air quality modeling for shanghai, China. *Transportation Research Record*, (2340), 38–48. https://doi.org/10.3141/2340-05

Mariescu-Istodor, R., & Fränti, P. (2018). CellNet: Inferring road networks from GPS trajectories. *ACM Transactions on Spatial Algorithms and Systems*, *4*(3). https://doi.org/10.1145/3234692

Marinelli, M., Palmisano, G., Astarita, V., Ottomanelli, M., & Dell'Orco, M. (2017). A Fuzzy set-based method to identify the car position in a road lane at intersections by smartphone GPS data. *Transportation Research Procedia*, *27*, 444–451. https://doi.org/10.1016/j.trpro.2017.12.047

Markovic, N., Sekula, P., Vander Laan, Z., Andrienko, G., & Andrienko, N. (2019). Applications of trajectory data from the perspective of a road transportation agency: Literature review and maryland case study. *IEEE Transactions on Intelligent Transportation Systems*, *20*(5), 1858–1869. https://doi.org/10.1109/TITS.2018.2843298

Markovic, N., Sekula, P., Vander Laan, Z., Andrienko, G., & Andrienko, N. (2019). Applications of trajectory data from the perspective of a road transportation agency: Literature review and maryland case study. *IEEE Transactions on Intelligent Transportation Systems*, *20*(5), 1858–1869. https://doi.org/10.1109/TITS.2018.2843298

Matte, C., Cunche, M., & Toubiana, V. (2018). *Does disabling Wi-Fi prevent my Android phone from sending Wi-Fi frames?* Retrieved from https://hal.inria.fr/hal-01575519v2

McCahill, C. (2017). Improving Last-Mile Connections to Transit: An Exploration of Data and Analysis Tools. *Congress for the New Urbanism*, 1–19. Retrieved from https://www.cnu.org/sites/default/files/2017_NewUrbanResearch_ImprovingLastMileConnectionstoTransit_McCahill.pdf

Mccormack, E., Scharnhorst, E., Zhao, W., & Tabat, D. (2011). *USING GPS TRUCK DATA TO IDENTIFY AND RANK BOTTLENECKS IN WASHINGTON STATE by*. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000444

Miller, H. J., O'Kelly, M. E., Jaegal, Y., Bachman, W., Huntsinger, L., & MacFarlane, G. (2017). *Estimating External Travel Using Purchased Third-Party Data*.

Montjoye, Y.-A., Hidalgo, C., Verleysen, M., & Blondel, V. (2013). Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific Reports*, *3*, 1376. https://doi.org/10.1038/srep01376

Monz, C., Mitrovich, M., D'Antonio, A., & Sisneros-Kidd, A. (2019). Using Mobile Device Data to Estimate Visitation in Parks and Protected Areas: An Example from the Nature Reserve of Orange County, California. *Journal of Park and Recreation Administration*, *37*(4), 92–109. https://doi.org/10.18666/jpra-2019-9899

Morrison S. (2020, July 8). *The hidden trackers in your phone, explained. How covert code enables your phone's apps to spy on you.* Vox. https://www.vox.com/recode/2020/7/8/21311533/sdks-tracking-data-location

NOAA. (2018, Sept. 27). *Selective Availability*. GPS.gov. https://www.gps.gov/systems/gps/modernization/sa/

O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A. F. gen., Penn, A., Fraser, D. S., & Jones, T. (2006). Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape. In P. Dourish & A. Friday (Eds.), *UbiComp 2006: Ubiquitous Computing* (pp. 315–332). Berlin, Heidelberg: Springer Berlin Heidelberg.

ODOT. (2019). *Oregon Commercial Truck Parking Study*.

Oosterlinck, D., Benoit, D. F., Baecke, P., & Van de Weghe, N. (2017). Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits. *Applied Geography*, *78*, 55–65. https://doi.org/10.1016/j.apgeog.2016.11.005

Pew Research Center. (2019, June 12). *Mobile Fact Sheet.* https://www.pewresearch.org/internet/fact-sheet/mobile/

Pinjari, A. R., Zanjani, A. B., Thakur, A., Irmania, A. N., Kamali, M., Short, J., … Park, L. (2014). Using Truck Fleet Data in Combination with Other Data Sources for Freight Modeling and Planning. *Pontificia Universidad Catolica Del Peru*, *8*(33), 44.

Porter, J. D., Kim, D. S., & Magana, M. E. (2011). *Wireless data collection system for real-time arterial travel time estimates*. Retrieved from http://www.oregon.gov/odot/td/tp_res/docs/reports/2011/wirelessdata.pdf

Qiu, J., & Wang, R. (2016). Inferring road maps from sparsely sampled GPS traces. *Journal of Location Based Services*, *10*(2), 111–124. https://doi.org/10.1080/17489725.2016.1183053

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, *33*(5), 727–748. https://doi.org/10.1068/b32047

Rescot, R. A. (2011). *Feasibility of Bluetooth Data as a Surrogate Analysis Measure of Traffic*. University of Kansas.

Roll, J. (2019). *Evaluating Streetlight Estimates of Annual Average Daily Traffic in Oregon*. Retrieved from https://trid.trb.org/View/1630252

Rosenworcel, J. (2019, May 15). correspondence to major US telecom companies. [Letter] AT&T. Federal Communications Commission. Retrieved from https://docs.fcc.gov/public/attachments/DOC-357494A2.pdf

Samsung. (2018, April 17). *What is the maximum range of a Bluetooth connection?* Samsung. https://www.samsung.com/levant/support/mobile-devices/what-is-the-maximum-range-of-a-bluetooth-connection/

Schlieder, C., & Yanenko, O. (2010). Spatio-temporal proximity and social distance: A confirmation framework for social reporting. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN 2010 - Held in Conjunction with ACM SIGSPATIAL GIS 2010*, 60–67. https://doi.org/10.1145/1867699.1867711

Sekuła, P., Marković, N., Vander Laan, Z., & Sadabadi, K. F. (2018). Estimating Historical Hourly Traffic Volumes via Machine Learning and Vehicle Probe Data: A Maryland Case Study. *Transportation Research Part C: Emerging Technologies*, *97*, 147–158. https://doi.org/10.1016/j.trc.2018.10.012

Sharifi, E., Hamedi, M., & Young, S. E. (2016). *Quality Assessment of Outsourced Probe Data on Signalized Arterials : Nine Case Studies in Mid-Atlantic*. 1–8.

Spinney, J. E. (2003). Mobile positioning and LBS applications. *Geography*, *88*(4), 256–265.

Statista. n.d. *Wireless subscriptions market share by carrier in the U.S. from 1st quarter 2011 to 2nd quarter 2020* Statista. https://www.statista.com/statistics/199359/market-share-of-wireless-carriers-in-the-us-by-subscriptions/

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30*(9), 1694–1716. https://doi.org/10.1080/13658816.2015.1099658

Stimpancic, J., Miranda-Moreno, L., & Labbe, A. (2016). *Measuring Congestion Using Large-Scale Smartphone-Collected GPS Data in an Urban Road Network*. https://doi.org/10.1080/19427867.2017.1374022

StreetLight Data. (2018). *Tools for Collecting Travel Behavior Data*. Retrieved from https://www.streetlightdata.com/wp-content/uploads/2018/03/EBook_Tools-for-Collecting-Travel-Behavior-Data.pdf

StreetLight Data. (2019). *StreetLight Multimode Methodology, Data Sources, and Validation*. (August), 36.

StreetLight Data. (2019). *StreetLight Data's Active Mode Metrics for Caltrans : Using Big Data to Analyze Bicycle and Pedestrian Trips Across the State of California Proprietary StreetLight Data RFP Response to ITB-DOT- Table of Contents*.

StreetLight Data. (2020). *StreetLight AADT 2019 Methodology and Validation White Paper*.

StreetLight Data. (2020, March, 29). *Count on Us for Counts: StreetLight Has Early 2019 Metrics.* Available from https://learn.streetlightdata.com/aadt2019-webinar

Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2012). Analyzing human activities through volunteered geographic information: Using flickr to analyze spatial and temporal pattern of tourist accommodation. *Lecture Notes in Geoinformation and Cartography*, (9783642342028), 57–69. https://doi.org/10.1007/978-3-642-34203-5_4

Tahlyan, D., Pinjari, A. R., Luong, T. D., & Ozkul, S. (2017). *Truck Route Choice Modeling using Large Streams of GPS Data*. Retrieved from http://cait.rutgers.edu/files/CAIT-UTC-NC-32-final.pdf%0Ahttps://cait.rutgers.edu/files/CAIT-UTC-NC-32-final.pdf%0Ahttps://rosap.ntl.bts.gov/view/dot/34888%0Ahttps://trid.trb.org/view/1504914

Tao, S., Manolopoulos, V., Rodriguez, S., & Rusu, A. (2012). Real-Time Urban Traffic State Estimation with A-GPS Mobile Phones as Probes. *Journal of Transportation Technologies*, *02*(01), 22–31. https://doi.org/10.4236/jtts.2012.21003

Thakur, A., Pinjari, A. R., Zanjani, A. B., Short, J., Mysore, V., & Tabatabaee, S. F. (2015). Development of algorithms to convert large streams of truck GPS data into truck trips. *Transportation Research Record*, *2529*, 66–73. https://doi.org/10.3141/2529-07

Thompson, S. A., Warzel C. (2019, Dec, 19) *Twelve Million Phones, One Data set, Zero Privacy.* The New York Times. https://www.nytimes.com/interactive/2019/12/19/opinion/location- tracking-cell-phone.html

Torrey, W. F., & Murray, D. (2017). *Managing Critical Truck Parking Tech Memo #2: Minnesota Case Study-Utilizing Truck GPS Data to Assess Parking Supply and Demand*. Retrieved from http://atri-online.org/wp-content/uploads/2017/02/Managing-Critical-Truck-Parking-Tech-Memo-2-02-2017-1.pdf

Tran, M. (2015). Accurate Location Detection. *911 Help SMS App*. Retrieved from https://transition.fcc.gov/pshs/911/Apps Wrkshp 2015/911_Help_SMS_WhitePaper0515.pdf

Transport for London. (2017). *Review of the TfL WiFi pilot*. Retrieved from http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf

Triggs, R. (2018, March 23). *A quick history of Bluetooth.* Android Authority. https://www.androidauthority.com/history-bluetooth-explained-846345/

Turner, S., & Koeneman, P. (2017). *Using Mobile Device Samples to Estimate Traffic Volumes*.

U.S. Department of Transportation. (2014, Nov. 13). *Satellite Navigation – GPS – Policy – Selective Availability.* Federal Aviation Administration. https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/techops/navservices/gnss/gps/policy/availability/

Utsch, P., & Liebig, T. (2012). Monitoring microscopic pedestrian mobility using bluetooth. *Proceedings - 8th International Conference on Intelligent Environments, IE 2012*, 173–177. https://doi.org/10.1109/IE.2012.32

Van Diggelen, Frank, Enge, Per, "The World's first GPS MOOC and Worldwide Laboratory using Smartphones," *Proceedings of the 28th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2015)*, Tampa, Florida, September 2015, pp. 361-369

Vasanthara, A., & Krishnamoorthy, K. (2016). Tire Pressure Monitoring System Using SoC and Low Power Design. *Circuits and Systems*, *07*(13), 4085–4097. https://doi.org/10.4236/cs.2016.713337

Weng, J., Liang, Q., Qiao, G., Chen, Z., & Rong, J. (2017). Taxi fuel consumption and emissions estimation model based on the reconstruction of driving trajectory. *Advances in Mechanical Engineering*, *9*(7), 1–12. https://doi.org/10.1177/1687814017708708

Xie, X., Wong, K. B. Y., Aghajan, H., Veelaert, P., & Philips, W. (2015). Inferring directed road networks from GPS traces by track alignment. *ISPRS International Journal of Geo-Information*, *4*(4), 2446–2471. https://doi.org/10.3390/ijgi4042446

Yan, Y., Schultz, M., & Zipf, A. (2019). An exploratory analysis of usability of Flickr tags for land use/land cover attribution. *Geo-Spatial Information Science*, *22*(1), 12–22. https://doi.org/10.1080/10095020.2018.1560044

Yang, F., Jin, P. J., Cheng, Y., Zhang, J., & Ran, B. (2015). Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. *International Journal of Sustainable Transportation*, *9*(8), 551–564. https://doi.org/10.1080/15568318.2013.826312

Yin, M., Sheehan, M., Feygin, S., Paiement, J. F., & Pozdnoukhov, A. (2017). A Generative Model of Urban Activities from Cellular Data. *IEEE Transactions on Intelligent Transportation Systems*, *19*(6), 1682–1696. https://doi.org/10.1109/TITS.2017.2695438

Young, S. E. (2012). *Bluetooth Traffic Detectors for Use as Permanently Installed Travel Time Instrument, State Highway Administration Research Report Project number SP909B4D*.

Young, S. E., Hou, Y., Sadabadi, K., Sekula, P., & Markow, D. (2017). Estimating Highway Volumes Using Vehicle Probe Data - Proof of Concept: Preprint. *ITS World Congress*, (March). Retrieved from https://www.osti.gov/servlets/purl/1426856

Yucel, S., Tuydes-Yaman, H., Altintasi, O., & Ozen, M. (2013). *DETERMINATION OF VEHICULAR TRAVEL PATTERNS IN AN URBAN LOCATION USING BLUETOOTH TECHNOLOGY*. 1–11.

Zalewski, A., Sonenklar, D., Cohen, A., Kressner, J., & Macfarlane, G. (2019). TCRP Synthesis of Transit Practice 138: Public Transit Rider Origin–Destination Survey Methods and Technologies. In *TCRP Synthesis of Transit Practice*. https://doi.org/10.17226/25428

Zhang, Y., Liu, J., Qian, X., Qiu, A., & Zhang, F. (2017). An automatic road network construction method using massive GPS trajectory data. *ISPRS International Journal of Geo-Information*, *6*(12). https://doi.org/10.3390/ijgi6120400

Zhao, D., Balusu, S. K., Sheela, P. V., Li, X., Pinjari, A. R., & Eluru, N. (2020). Weight-categorized truck flow estimation: A data-fusion approach and a Florida case study. *Transportation Research Part E: Logistics and Transportation Review*, *136*(April 2019). https://doi.org/10.1016/j.tre.2020.101890

Zinner, S. (2012). *A Methodology for Using Bluetooth To Measure Real-Time Work Zone Travel Time*. Georgia Institute of Technology.